**UNIVERSIDAD AUTÓNOMA DE MADRID**
**ESCUELA POLITÉCNICA SUPERIOR**


**Grado en Ingeniería Informática**

# TRABAJO FIN DE GRADO

## Análisis de factores discriminantes y algoritmos para la detección de reincidencia en casos de violencia de género.

**Autor: Guillermo Rodríguez Lorenzo**
**Tutor: Lara Quijano Sánchez**
**Ponente: Iván Cantador Gutiérrez**

**junio 2020**

**Guillermo Rodríguez Lorenzo**
***Análisis de factores discriminantes y algoritmos para la detección de reincidencia en casos de violencia de género.***

**Guillermo Rodríguez Lorenzo**

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

# Resumen

La violencia de género es un problema que afecta a millones de personas en todo el mundo. Para combatir tal problema, en España existe un protocolo de seguimiento policial que tiene como objetivo reducir la probabilidad de reincidencia en los casos en los que se ha producido algún tipo de violencia de género. Este protocolo se complementa con VioGén, un sistema creado por la Secretaría de Estado de Seguridad del Ministerio del Interior para la predicción del riesgo. El objetivo de VioGén es facilitar la labor de la autoridad competente a la hora de decidir las medidas de seguridad y protección más apropiadas para cada caso. La última versión del sistema está en uso desde 2017. En este trabajo se ha estudiado el actual sistema de predicción y una base de datos anonimizada extraída del actual sistema en vigor y facilitada por la Secretaría de Estado de Seguridad. Dicha base de datos contiene la evolución de los 44.655 casos de violencia de género registrados entre octubre de 2016 y marzo de 2019. Mediante el uso de técnicas de Ciencia de Datos y Aprendizaje Automático se han desarrollado nuevos modelos de predicción alternativos al actual sistema VioGén, para su estudio. Este documento recoge el proceso de limpieza, análisis y codificación que se ha realizado sobre la base de datos proporcionada, los diseños y la metodología empleada en cada uno de los nuevos modelos y la evaluación de los resultados obtenidos, comparándose dichos resultados con los proporcionados por el sistema de predicción actual.

# Palabras clave

Aprendizaje Automático, Ciencia de Datos, Violencia de Género.

# Abstract

Gender violence is a problem that concerns millions of people around the world. To fight this problem in Spain, a police monitoring protocol has been created to reduce the likelihood of relapse in cases where there has been some kind of gender violence. This protocol is complemented by VioGén, a system created by the Secretary of State for Security of the Ministry of the Interior for risk prediction. VioGén's goal is to help the authorities decide what are the most appropriate security and protection measures for each case. The latest version of the system has been in use since 2017. In this work, the current forecasting system has been studied along an anonymized database provided by the Secretary of State for Security and extracted from the current system. This database contains the evolution of the 44.655 cases of gender violence registered between october 2016 and march 2019. Through the use of Data Science and Machine Learning techniques, new prediction models have been studied as an alternative to the current VioGén system. This document includes the cleaning, analysis and coding process that has been carried out on the database provided, the designs and the methodology used in each of the new models and the evaluation of the results obtained, comparing these results with those provided by the current prediction system.

# Keywords

Machine Learning, Data Science, Gender Violence.

# ÍNDICE

# Listas

## Lista de ecuaciones



## Lista de figuras

# Lista de tablas

# 1

# Introducción

Desafortunadamente, la violencia de género es una de las violaciones de los derechos humanos más común y afecta a millones de personas [1]. Constituye un atentado contra la libertad, integridad y dignidad de la víctima. La violencia de género, comprende todo acto de violencia física y psicológica, incluidas las agresiones a la libertad sexual, las amenazas, las coacciones o la privación arbitraria de libertad [2]. Para combatir tal problema, en España existe un protocolo de seguimiento policial, que tiene como objetivo reducir la probabilidad de reincidencia en los casos en los que se ha producido algún tipo de violencia de género. Este protocolo se complementa con VioGén, un sistema creado por la Secretaría de Estado de Seguridad del Ministerio del Interior (SES) para la predicción del riesgo. El objetivo de VioGén es facilitar la labor de la autoridad competente (AC) a la hora de decidir las medidas de protección más apropiadas para cada caso. La última versión del sistema está en uso desde 2017.

En este trabajo se propondrán alternativas al actual sistema de predicción. Para ello, se utilizará una base de datos anonimizada extraída del actual sistema en vigor y facilitada por la SES. Dicha base de datos contiene la evolución de 44.655 casos de violencia de género registrados entre octubre de 2016 y marzo de 2019. Mediante el uso de técnicas de Ciencia de Datos y Aprendizaje Automático se extraerá y procesará la información de la base de datos, y se desarrollarán y estudiarán a partir de esta nuevos modelos de predicción.

## 1.1. Motivación

Hasta ahora, la construcción del actual sistema de predicción [3] se ha abordado desde el área de las ciencias sociales, donde se ha analizado la importancia de cada indicador en un conjunto pequeño de casos y se ha realizado su correspondiente validación. En la actualidad, las herramientas de Ciencia de Datos y Aprendizaje Automático se emplean con éxito en numerosos ámbitos para transformar cantidades voluminosas de datos en conocimiento. Dichas herramientas se usan, por ejemplo, para predecir la probabilidad de lesión de un deportista [4] o para detectar clientes insatisfechos [5]. En el campo de la seguridad se han utilizado estas herramientas para predecir posibles actividades criminales (*Predictive policing*) [6]. En este trabajo se pretende usar el potencial de estas herramientas para,

a partir de un conjunto más amplio de datos, proponer nuevos modelos de predicción del riesgo.

## 1.2. Objetivos y tareas

Los objetivos principales de este TFG son dos: i) Analizar los datos facilitados, estudiando posibles factores discriminantes y predictivos; ii) Estudiar modelos alternativos de predicción del riesgo, que empleen técnicas de Aprendizaje Automático. Cabe destacar que hasta ahora solo se había estudiado si el agresor iba a reincidir, observando una foto estática del problema y definiendo la reincidencia (como se detallará más adelante) de forma genérica, a partir de diferentes indicadores de violencia y quebrantamientos. En este TFG se ampliará este enfoque de tres maneras: i) Estudiando distintos tipos de reincidencia; ii) Estudiando la relevancia de variables exógenas que no se estén considerando; iii) Estudiando variables predicativas que reflejen toda la evolución del caso. Otro factor a analizar es que el modelo actual de VioGén, a la hora de realizar predicciones de riesgo en casos ya existentes, no tiene en cuenta el nivel de protección que se le asignó a la víctima anteriormente. Es decir, no distingue si no ha habido reincidencia porque el nivel de protección ha sido muy alto o, si por el contrario, ha sido porque la situación realmente carecía de peligro. En este trabajo intentaremos tener en cuenta este factor a la hora de diseñar los modelos.

Para alcanzar los objetivo propuestos, se seguirá un pipeline de Ciencia de Datos [7], realizándose las siguientes tareas:

- Limpieza de la nueva base de datos.
- Estudio de los datos y de las variables a utilizar.
- Inclusión de variables exógenas e historial.
- Transformación y codificación de los datos a entradas para los modelos.
- Estudio de las variables a predecir.
- Creación y análisis de modelos.

## 1.3. Estructura del documento

Con respecto a la estructura de este documento, en el capítulo 2 se presenta el estado del arte del actual sistema y otros sistemas existentes, así como las técnicas de Aprendizaje Automático utilizadas. En el capítulo 3 se muestra el proceso de limpieza y análisis de la base de datos. Seguidamente, se explican en el capítulo 4 los modelos de predicción diseñados. A continuación, en el capítulo 5 se evalúan los resultados obtenidos. Se concluye con el capítulo 6, en el que se proponen líneas futuras de trabajo.

# 2

# Estado del arte

## 2.1. Sistema actual

El sistema VioGén (cuya última versión, 4.0, fue instaurada en el año 2017), contiene dos herramientas: el formulario de Valoración Policial del Riesgo (VPR) y el formulario de Valoración Policial de la Evolución del Riesgo (VPER) . Estos formularios se muestran en detalle en el apéndice A.

El protocolo seguido en casos de violencia de género, que incluye la utilización del sistema VioGén, es el siguiente: cuando la víctima acude por primera vez a denunciar los hechos a la institución pertinente, la AC rellena el formulario VPR complementando las repuestas que haya facilitado la víctima con sus propias investigaciones. Los resultados de ese formulario se ejecutan en el actual modelo de predicción de riesgo y el sistema devuelve una recomendación del nivel de riesgo a asignar. Este nivel puede ser no apreciado, bajo, medio, alto o extremo. La  decide posteriormente cuál es el riesgo que se le asigna a la víctima. El riesgo asignado conlleva una serie de medidas de protección (recomendadas a seguir) y, además, establece un plazo máximo en el que se deberá realizar un revisión del caso, consultando nuevamente a la víctima [3] . La tabla 2.1 muestra el plazo máximo correspondiente a cada nivel.

| **Nivel de Protección** | **Plazo máximo hasta la siguiente revisión.** |
|---|---|
| Extremo | 72 horas |
| Alto | 7 días |
| Medio | 30 días |
| Bajo | 60 días |
| No apreciado | 60 días, solo si existe una orden de protección en vigor |

**Tabla 2.1:** Plazo máximo para realizar la siguiente revisión

A partir de este momento, cada vez que la víctima acuda a realizar una de las revisiones periódicas, la AC rellenará un formulario VPER. De forma análoga al caso anterior, los resultados son introducidos en un segundo modelo de predicción, (generado esta vez a partir de las respuestas recogidas en los formularios VPER), que actualizará el riesgo recomendado. La AC volverá a reevaluar después cuál

es el nivel de protección más apropiado, modificando las medidas de seguridad en caso necesario y estableciendo el plazo en el que se tiene que realizar la siguiente revisión de seguimiento periódica, en función del nivel de protección actualizado, utilizándose los mismos plazos expuestos anteriormente. Nótese que la víctima puede acudir a denunciar nuevos hechos que se hayan producido antes de la siguiente revisión periódica. Si esto ocurre, al igual que en el caso anterior, se rellenará un formulario VPER con la nueva información recopilada y se reevaluará el nivel de riesgo, modificándose las medidas de seguridad en caso necesario y estableciéndose el nuevo plazo para la siguiente revisión periódica. En consecuencia, dentro del formulario VPER existen dos modelos: modelo VPER con reincidencia y modelo VPER sin reincidencia. El modelo VPER sin reincidencia se rellena en cada una de las revisiones periódicas a la que acude la víctima cuando no se observa que haya habido reincidencia. En este modelo las respuestas a los cuatro primeros bloques de preguntas están predefinidas como "vacías" para agilizar el proceso. El modelo VPER con reincidencia se rellena, o bien en la misma situación anterior, cuando **sí** que se observa que ha habido reincidencia, o bien cuando la víctima acude a denunciar nuevos hechos antes de la siguiente revisión periódica.

El modelo de predicción del riesgo [3] que se encuentra en vigor desde 2017, tiene un enfoque estadístico. Para su construcción, en primer lugar, se calcula el peso de cada indicador, que será el *Odds Ratio*(OR) del propio indicador con respecto a la reincidencia durante seis meses. A continuación, para cada caso se obtiene un valor numérico del riesgo, sumando los pesos de los indicadores presentes en el caso. Después se ordenan los casos en orden decreciente, con respecto al valor numérico del riesgo, obteniendo una escala empírica. Por último, sobre esa escala se establecen cinco intervalos, cada uno asociado a un nivel de riesgo. Los cortes se determinan viendo cuántos reincidentes y no reincidentes quedan en cada tramo. Para cada nuevo caso se calcula el valor numérico del riesgo sumando los pesos de los indicadores presentes. En función del intervalo donde se encuentre este valor se determinará el nivel de riesgo. Los detalles específicos sobre el funcionamiento del modelo son difíciles de conocer, debido a la sensibilidad de los datos y a que tanto el modelo como la tesis [3] están protegidos por su carácter sensible.

## 2.2. Otras herramientas

Existen otras herramientas y aplicaciones relacionadas con la violencia de género, que han sido creadas por diferentes instituciones con el fin de informar y proteger a la víctimas. Algunas de ellas son: "LIBRES" [8], "YgualeX" [9], "AlertCops" [10], "PORMI" [11], "Trusted Contacts" [12] y "Seguras" [13]. No obstante, estudiando sus características observamos que ninguna de ellas cuenta con un algoritmo de estimación del riesgo de reincidencia.

## 2.3. Técnicas empleadas para la creación, implementación y validación de modelos.

En esta sección se pretende introducir la base teórica de las técnicas que se han utilizado a la hora de crear, implementar y evaluar los modelos de predicción, así como explicar la metodología que se ha empleado. Los modelos se han implementado en el lenguaje de programación *Python*, usando principalmente las librerias *Scikit-learn, Numpy* y *Pandas*.

La metodología empleada es común a todos los modelos, con independencia del diseño. En primer lugar, se elige la métrica que se va a utilizar para la evaluación de los resultados. A continuación, se balancea el conjunto de datos de entrenamiento con respecto a la variable a predecir y se eliminan las variables predictoras que no son significativas. Seguidamente, se valida el modelo con cada uno de los algoritmos de predicción, buscando para cada algoritmo los hiperparámetros que mejor se ajustan al modelo. Por último, se escoge para cada modelo el algoritmo que mejores resultados haya obtenido, o se elige una combinación de varios algoritmos.

### 2.3.1. Técnicas con respecto a la creación de modelos

#### Algoritmos de predicción

En cada uno de los modelos de predicción creados, el pilar principal que determina su funcionamiento es el algoritmo de aprendizaje automático utilizado. Dentro de los algoritmos de aprendizaje automático, se han utilizado algoritmos de aprendizaje supervisado. Estos últimos son aquellos que realizan sus predicciones a partir de un conjunto de datos para los que la respuesta es conocida. Los algoritmos de aprendizaje supervisado se pueden dividir en algoritmos de clasificación, en los que la respuesta a predecir es una categoría y, algoritmos de regresión, en los que la respuesta a predecir es un valor continuo [14]. No obstante, partiendo de algoritmos de regresión, también podemos clasificar en categorías, en función del resultado obtenido. Por ejemplo, aunque queramos predecir si en un caso va a haber reincidencia o no, fenómeno que se puede codificar con una variable binaria, se pueden utilizar algoritmos de regresión en los que las predicciones sean valores en el intervalo [0,1] e, interpretar estos valores como la probabilidad de reincidencia.

De manera general a todos los algoritmos, cada entrada de nuestro conjunto de datos se puede dividir en variables predictoras $(X_1, \ldots, X_n)$ y variable respuesta $Y$, que es la que queremos predecir. Dada una nueva entrada $(X_{01}, \ldots, X_{0n})$, los algoritmos utilizan el conjunto de datos de entrenamiento de tamaño $M$ , $\{(X_{i1}, \ldots, X_{in}, Y_i) = (X_i, Y_i)$, con $1 \leq i \leq M\}$, para predecir la variable respuesta $Y_0$.

Los algoritmos de aprendizaje supervisado que se han utilizado se describen a continuación:

**Algoritmo k vecinos más próximos (k-NN)**: Dada una nueva entrada $X_0$ se seleccionan

las $k$ entradas del conjunto de entrenamiento cuyas variables predictoras $X_{i_1}, \ldots, X_{i_k}$ tengan menor distancia a la variable $X_0$. Se predice la variable respuesta $Y_0$ en función de las respuestas $Y_{i_1}, \ldots, Y_{i_k}$. Hay que tener en cuenta que para la elección del algoritmo, es necesario determinar la métrica para medir la distancia (Euclídea, Minkowski, etc), el número $k$ de vecinos y cómo influyen las variables $Y_{i_1}, \ldots, Y_{i_k}$ en la respuesta $Y_0$ (influencia uniforme, influencia ponderada, etc).

**Algoritmo regresión lineal**: En este caso, dada una nueva entrada $X_0$, la predicción de la variable respuesta será $\hat{Y_0} = \beta_0 + \beta_1 X_{01} + ... + \beta_n X_{0n}$. Dado el conjunto de datos de entrenamiento, los parámetros $\beta_0, \beta_1, ..., \beta_n$ escogidos son aquellos que minimicen el error de mínimos cuadrados $\sum_{i=1}^{M}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_n X_{in}))^2$.

**Algoritmo de regresión logística binario**: El algoritmo de regresión logistica se usa en los casos en los que la variable respuesta se puede codificar de forma binaria (0 o 1). Dada la entrada $X_0$, definimos:

$$P(Y_0 = 1|X_0) = p_0 = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{01} + ... + \beta_n X_{0n})}}. \tag{2.1}$$

En caso de querer clasificar la variable respuesta entre dos categorías, si la probabilidad es mayor que 0,5, lo clasificaremos dentro de la categoría 1 y, dentro de la categoría 0, en caso contrario. A partir del conjunto de datos de entrenamiento, los parámetros $\beta_0, \beta_1, ..., \beta_n$ escogidos son aquellos que maximicen la función de verosimilitud $\prod_{i=1}^{M} p_i^{Y_i}(1-p_i)^{1-Y_i}$. Este algoritmo se puede extender a casos en los que la variable respuesta puede pertenecer a más de dos categorías, denominándose, en este caso, regresión logística multinomial [15].

**Algoritmos Naive-Bayes**: Estos algoritmos clasifican la variable respuesta en una de las $k$ posibles categorías $C_1, ..., C_k$. Dada una nueva entrada $X_0$, se calculan las probabilidades condicionadas $P(C_1|X_0), ..., P(C_k|X_0)$ y se elige la clase que tenga mayor probabilidad de ocurrencia. Sin embargo, el cálculo de estas probabilidades es complejo, por lo que resulta de gran utilidad aplicar la regla de Bayes:

$$P(C_i|X_0) = \frac{P(C_i)P(X_0|C_i)}{P(X_0)}. \tag{2.2}$$

Adicionalmente, el denominador se podrá suprimir, dado que su cálculo no depende de $C_i$ y la entrada $X_0$ será dada. Por lo tanto, el cálculo se reduce a encontrar la categoría $i$, que maximice $P(C_i)P(x_0|C_i)$. En el caso de que todas las categorías tengan la misma probabilidad de ocurrencia, únicamente habrá que calcular $P(X_0|C_i)$. Podemos encontrar distintos algoritmos en función de la distribución que se considera que se sigue. En Naive-Bayes Bernoulli, si las variables predictoras $X_{1*}$,...$X_{n*}$ son binarias, para cada categoría i se considera que las variables $X_{1*}|C_i, ..., X_{n*}|C_i$ son independientes y siguen una distribución de Bernoulli. En Naive-Bayes Multinomial, si $X_*$ es el vector de variables predictoras, se considera que para cada categoría $i$ el vector $X_*|C_i$ sigue una distribución multinomial. Dependiendo del conjunto de variables predictoras que tengamos, es posible que sea ne-

cesario binarizarlas para adecuarlo al modelo Naive-Bayes que vayamos a usar.

**Algortimos SVM**: son algoritmos para clasificación entre dos categorías. La idea básica es considerar cada entrada del conjunto de entrenamiento $X_{i1}, ..., X_{1n}$ como un punto en un espacio $n$-dimensional e intentar separar el espacio en dos porciones mediante un hiperplano, de tal manera que las entradas situadas en la primera porción se considera que pertenecen a una de las categorías y las situadas en la segunda a la otra. Se busca el hiperplano óptimo que minimice el número de puntos que se encuentren en la porción del espacio errónea. Cuando tengamos una nueva entrada $X_0$, se le asignará la categoría correspondiente a la porción del espacio donde quede situada. Este algoritmo SVM se conoce como lineal. Sin embargo, no siempre es posible separar el espacio de manera razonable con un hiperplano. En tales casos, se pueden utilizar funciones que transformen el espacio en otros en los que sí sea posible utilizar hiperplanos para realizar la separación. Estas funciones se denominan funciones $kernel$. En el caso de clasificaciones en las que se puede elegir entre $n$ categorías con $2 < n$, una de las opciones es crear $n(n-1)/2$ clasificadores y escoger la categoría que haya sido elegida un mayor número de veces.

**Algoritmo Random Forest**: El algoritmo Random Forest combina el resultado de $m$ árboles de decisión. Cada uno de ellos se crea a partir de un subconjunto de tamaño $k$ del conjunto de datos de entrenamiento, elegido aleatoriamente con reemplazamiento. En cada uno de los árboles solo se tendrán en cuenta $t$ variables de las $n$ variables predictoras. Para general los árboles de decisión se puede utilizar el algoritmo ID3 [16].

### Obtención de hiperparámetros

Como se ha podido observar a medida que se presentaban los algoritmos, cada uno de ellos tiene una lista de parámetros que se deben establecer antes de comenzar el proceso de aprendizaje. Ejemplos de estos parámetros son el número de vecinos en el algoritmo k-NN o la función $kernel$ en el algoritmo SVM. Lo que haremos en cada caso será entrenar sucesivas veces el algoritmo y evaluarlo, variando los valores de los parámetros dentro de unos intervalos razonables. Nos quedaremos con los parámetros que mejor resultado den según la métrica de evaluación que establezcamos. Esta técnica se conoce como *Hyperparameter Tuning*.

### Combinación de algoritmos

Una vez seleccionados los algoritmos que mejor funcionan, se puede generar un nuevo meta-algoritmo que combine los resultados de los anteriores. Una de las técnicas disponibles para ello es el método de combinación *Stacking*. Con carácter general, este método utiliza dos conjuntos de entrenamiento ($E_1$ y $E_2$), $m$ algoritmos débiles y un algoritmo principal. Cada uno de los algoritmos débiles se entrena con el conjunto de entrenamiento $E_1$. A continuación, se genera un nuevo conjunto de entrenamiento $\hat{E}_2$, a partir del conjunto $E_2$. Más concretamente, para cada entrada $(X_i, Y_i)$ del

conjunto $E_2$, se predice a partir de $X_i$ la variable respuesta $Y_i$ con cada uno de los $m$ algoritmos débiles, obteniendo $(\hat{Y}_{i1}, ...\hat{Y}_{im}, Y_i)$. El algoritmo principal se entrena con el conjunto $\hat{E}_2$. Cuando llegue una nueva entrada $X_0$, cada uno de los $m$ algoritmos débiles nos dará una predicción $\hat{Y}_{01}, ...\hat{Y}_{0m}$. Estas predicciones se introducirán en el algoritmo principal obteniendo la predicción final de $Y_0$. Otra opción es combinar las predicciones de los $m$ algoritmos débiles escogiendo la media de las predicciones (*Avering*) o la predicción que más veces haya salido (*Voting*), en lugar de utilizar un algoritmo principal [17].

### 2.3.2. Técnicas con respecto a la preparación del conjunto de datos de entrenamiento.

#### Balanceamiento de carga

Uno de los problemas más frecuentes a la hora de generar los modelos de predicción es que nuestro conjunto de datos de entrenamiento no esté balanceado. Es decir, en nuestro conjunto de datos de entrenamiento aparecen muchas más veces unas categorías que otras. En esta situación, los algoritmos siempre predicen las categorías mayoritarias obviando el resto de categorías. Por ejemplo, en nuestro caso particular, hay muchos más supuestos en los que no se produce reincidencia que en los que sí. Si generásemos los algoritmos con este conjunto de datos, tal cual, estos siempre predirán que no va a haber reincidencia. Para solucionar este problema, lo que vamos a hacer es balancear los datos. Suponemos que $C_1, C_2, ..., C_m$ son las posibles categorías de la variable respuesta ordenadas en orden creciente con respecto al número de veces que aparecen en nuestro conjunto de datos y $l$ el número de veces que la variable respuesta pertenece a la categoría $C_1$. Seleccionaremos para el conjunto de datos entrenamiento todos las entradas cuya variable respuesta sea $C_1$. Para cada una de las otras $m-1$ categorías añadiremos al conjunto de entrenamiento $l$ entradas seleccionadas aleatoriamente sin remplazamiento de entre las que la variable respuesta pertenezca a esa categoría.

#### Elección de variables.

Por otra parte, aunque tengamos un gran número de variables predictoras, puede haber variables que no sean representativas o aporten ruido al modelo. Para eliminar estas variables vamos a utilizar el método de regularización de Lasso [18]. Este método se basa en minimizar el error de mínimos cuadrados introduciendo un factor de penalización $\lambda$ para intentar reducir el número de variables. En concreto, dado el conjunto de datos de entrenamiento de tamaño $M$, $\{(X_{i1}, \ldots, X_{in}, Y_i)$, con $1 \leq i \leq M\}$, el error a minimizar será $\frac{1}{2M}\sum_{i=1}^{M}(Y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_n X_{in}))^2 + \lambda \sum_{j=1}^{n} |\beta_j|$. Una vez calculados los coeficientes $\beta_0, \beta_1, ..., \beta_n$ que minimizan el error, se eliminan aquellas variables cuyos coeficientes asociados sean 0. Adicionalmente, podemos utilizar los propios coeficientes $\beta_0, \beta_1, ..., \beta_n$ obtenidos para, dada una nueva entrada $(X_{01}, \ldots, X_{0n})$, predecir la variable respuesta $Y$ como $\hat{Y}_0 =$

$\beta_0 + \beta_1 X_{01} + ... + \beta_n X_{0n}$. Este algoritmo de predicción se conoce como regresión Lasso y es una variante del algoritmo de regresión lineal. Cabe destacar que Lasso elimina las variables que no son representativas al modelar el problema mediante regresión lineal. Sin embargo, podría ocurrir que al modelar el problema con otros algoritmos sí fuesen relevantes. Por lo tanto, conviene verificar que las variables eliminadas por Lasso son variables de baja activación o variables con poca correlación con la respuesta. Otras opciones para la eliminación de variables serían el análisis de componentes principales (PCA) [19] o la regresión de mínimos cuadrados parciales (PLS) [20], así como técnicas ad hoc concretas para optimizar parámetros dentro de cada algoritmo de clasificación utilizado. Siendo conscientes de que estas otras técnicas podrían mejorar los resultados presentados en este trabajo, se deja como trabajo futuro el estudio de métodos alternativos y ad hoc de selección de variables.

### 2.3.3. Técnica para evaluación de modelos.

#### Validación Cruzada

Para evaluar los modelos propuestos ulilizaremos validación cruzada de $k$ iteracciones (*k-cross validation*). Utilizamos esta técnica dado que garantiza mayor precisión que destinar directamente parte de nuestro conjunto de datos para entrenamiento y parte para validación.
En la validación cruzada de $k$ iteraciones se dividen las entradas de nuestro conjunto de datos $P$ en k subconjuntos $P_1, P_2, ..., P_k$ de tamaño lo más parecido posible. En cada iteracción $j$, se utiliza el conjunto de datos $P \setminus P_j$ para entrenar el modelo y el conjunto de datos $P_j$ para validarlo. Al final de las $k$ iteracciones tendremos una predicción para cada entrada de nuestro conjunto de datos.
En nuestro caso particular, si $M$ es el número de entradas de nuestro conjunto de datos, utilizaremos validación cruzada con $k = M$ iteraciones (*leave-one-out cross validation*), salvo si estamos buscando los mejores hiperparámetros de un algoritmo. En este último caso, utilizaremos $k = \lfloor \frac{M}{10} \rfloor$ iteracciones para mayor rapidez.

#### Métrica de evaluación de modelos

Para seleccionar el mejor modelo, es necesario definir una métrica que nos indique cómo de fiables son las predicciones hechas por cada modelo. Cuando la variable a predecir sea binaria, una opción es escoger el modelo cuya puntuación *F1* sea mayor. Para entender el funcionamiento de esta métrica es necesario introducir los siguientes conceptos previos:
**Verdaderos positivos (TP)**: casos en los que tanto la clase real como la predicha es 1 (verdadero).
**Falsos positivos (FP)**: casos en los que clase real es 0 (falso) y la predicha es 1 (verdadero).
**Verdaderos negativos (TN)**: casos en los que tanto la clase real como la predicha es 0 (falso).
**Falsos negativos (FN)**: casos en los que la clase real es 1 (verdadero) y la predicha es 0 (falso).
**Sensibilidad (Recall)**: porcentaje de casos identificados correctamente de entre aquellos en los que la clase real es 1 (verdadero): $\frac{TP}{TP+FN}$

**Precisión**: porcentaje de casos identificados correctamente como positivos del total del número de casos identificados como positivos : $\frac{TP}{TP+FP}$.

La puntuación F1 combina las métricas de sensibilidad y precisión, facilitando un solo valor con el que poder medir cómo de fiable es el modelo:

$$F1 = \frac{2 \times sensibilidad \times precisión}{sensibilidad + precisión} \tag{2.3}$$

La puntuación $F1$ es una media armónica de la precisión y la sensibilidad, de tal manera que, si la diferencia entre las dos medidas anteriores es grande, será más representativa en la puntuación $F1$ la medida que sea más pequeña. Nos quedaremos con el modelo cuya puntuación $F1$ sea mayor.

En el caso de clasificaciones en las que hay más de dos categorías, se pueden extender las nociones de precisión y sensibilidad [21] y utilizar la puntuación $F1$. No obstante, en nuestro caso, utilizaremos una métrica específica que se explicará más adelante (véase Sección 4.1).

# 3

# Limpieza y análisis de la base de datos

## 3.1. Introducción

A partir de la base de datos proporcionada por la SES, que se describe en detalle a continuación, se pretende obtener dos conjuntos de datos, que posteriormente serán utilizados para generar los modelos de predicción.

En el primer conjunto cada entrada estará ligada a un formulario VPR y contendrá las respuestas de este, así como variables sociodemográficas genéricas del caso, como pueden ser la edad del autor, la edad de la víctima o la localización del caso. La entrada también tiene asociada el nivel de protección que finalmente se aplicó (NPA). Se utilizará este primer conjunto para crear modelos que predigan el riesgo cuando se produzca un nuevo caso. En el segundo conjunto de datos cada entrada estará ligada a un formulario VPER y contendrá las respuestas de este, las variables sociodemográficas génericas (que son constantes en un caso y se heredan del anterior conjunto de datos), el NPA tras ese formulario y, de forma novedosa, cada entrada contendrá información sobre cómo ha ido evolucionando el caso hasta el momento en el que se rellena el formulario. Dicha información se conoce como el histórico del caso y se explica con mayor detalle en el apartado 3.6. Este segundo conjunto se utilizará para crear modelos que actualicen el riesgo en cada revisión del caso (que como se vio anteriormente pueden ser por una nueva denuncia o por una revisión programada).

## 3.2. Base de datos inicial

La base de datos proporcionada por la SES se compone de 8 tablas: *Autores* (10.612 registros), *Denuncias* (13.370 registros), *Víctimas* (10.612 registros), *Casos* (10.622 registros), *Hechos* (14.561 registros), *Histórico* (10.911 registros), *Formularios VPR* (46.047 registros) y *Formularios VPER* (255.425 registros). Los campos que contienen las tablas proporcionadas se describen de manera detallada en el anexo B. Adicionalmente, se han creado dos nuevas tablas, *PoblacionLocalidad* y *PoblacionProvincia*, que recogen respectivamente, el número de habitantes de cada municipio y cada provincia del territorio español. Estas dos últimas tablas se han creado a partir de los datos extraídos

de [22]. La finalidad de añadir estas dos tablas es poder conocer el número de habitantes del municipio y la provincia donde se han producido los hechos e incluir dichos datos en la información general asociada al caso, por si fuese relevante. A continuación se detallan los atributos extraídos de las tablas proporcionadas por la SES:

**Tabla formularios VPER:** se extraerán las respuestas que se recogieron al rellenar los formularios VPER, la fecha en la que se registraron estos y el nivel de protección que se asignó a partir de los mismos. Los formularios VPER (como se puede apreciar en el Anexo A) están formados por 7 tipos distintos de preguntas:

**Preguntas tipo A**: en las que únicamente se puede responder “Sí” o “No”.

**Preguntas tipo B**: en las que se puede responder “Sí”, “No” o “No Sabe”.

**Preguntas tipo C**: en las que se puede responder “Sí”, “No” o “No procede”.

**Preguntas tipo D**: en las que se puede contestar “Leve”, “Grave” o “Muy grave”.

**Preguntas tipo E**: en las que se puede elegir más de una opción (respuestas multiopción).

**Preguntas tipo F**: en las que se puede contestar “Nulo”, “Bajo” o “Alto”.

**Preguntas tipo G**: en las que se puede contestar “Infravalora”, “Sobrevalora” o “Igual”.

**Tabla formularios VPR:** se extraerán las respuestas que se recogieron al rellenar los formularios VPR de cada caso, la fecha en la que se registraron estos y el nivel de protección que se asigno a partir de los mismos. El formulario VPR contiene preguntas de todos los tipos presentes en los formularios VPER, salvo preguntas de tipo C, F y G.

**Tabla Autor y Tabla Víctima :** estas tablas aportan datos anonimizados sobre los autores y víctimas asociadas a cada caso. De entre toda la información recogida, únicamente utilizaremos la fecha de nacimiento del autor y de la víctima, ya que el resto de registros tiene una incidencia inferior al 25 %. A partir de estas fechas y de la fecha de registro del formulario VPR del caso, se obtendrá la edad del autor y de la víctima en el momento que se produjeron los hechos.

**Tabla Casos:** se extraerá la institución donde se ha denunciado el caso.

**Tabla denuncias:** se extraerá la provincia donde se ha denunciado el caso. Dicho dato servirá para posteriormente conocer la población de la provincia donde se ha producido el caso, a partir de la combinación con la tabla *poblacionProvincia*.

**Tabla Histórico:** de esta tabla recopilaremos información sobre los intervalos de tiempo en los que los casos están inactivos. Dicha información se utilizará únicamente con fines estadísticos. No obstante, podría ser incluida en los modelos en futuros trabajos.

**Tabla Hechos:** se extraerá la localidad donde se han producido los hechos. Dicho dato servirá para posteriormente conocer la población de la localidad donde se ha producido el caso, a partir de la combinación con la tabla *poblacionLocalidad*.

## 3.3. Limpieza de la base de datos

Una vez extraídos los datos de las tablas, ha sido necesario efectuar una limpieza para solventar los errores e incoherencias presentes y completar o corregir campos de la base de datos que estuviesen vacíos.

### 3.3.1. Corrección y establecimiento de coherencia en las respuestas a los formularios VPR y VPER

Comenzamos completando todos los campos vacíos en las respuestas a los formularios VPR y VPER, y corrigiendo las incoherencias entre preguntas anidadas. De manera general, cuando la respuesta a una de las preguntas esta vacía se toma como respuesta “No”, si la pregunta es de tipo A, “No sabe”, si es de tipo B, “No procede”, si es de tipo C, “Bajo”, si es de tipo F e “Igual”, si es de tipo G.

Seguidamente, pasamos a corregir incoherencias entre las respuestas dadas en preguntas anidadas. A modo de ejemplo, se dan casos en los que se registra que la víctima no ha recibido agresiones, mientras que la subpregunta relacionada con el nivel de las agresiones refleja que ha sido alto. En los formularios VPR y VPER existen preguntas de tipo A y B que tienen anidadas en primer nivel al menos una pregunta de tipo A, B, D o E, como se puede ver en el apéndice A. El procedimiento para la corrección de estos errores es el siguiente:

- 1.Si la respuesta a alguna de las preguntas contenidas de tipo A o B es “Sí” o la respuesta a alguna de las preguntas contenidas de tipo D está rellena o la respuesta a alguna de las preguntas contenidas de tipo E tiene marcada alguna de las multiopciones, se tomará como respuesta a la pregunta contenedora “Sí”.
- 2. Si la respuesta a la pregunta contenedora es “Si”, las respuestas a preguntas contenidas de tipo D sin responder se tomarán como "Leve".
- 3. Si después de revisar los puntos anteriores, la respuesta a la pregunta contenedora sigue siendo “No”, todas las respuestas a preguntas contenidas de tipo B pasan a ser “No”. Es decir, no se permite que la respuesta contenedora sea “No” y haya respuestas contenidas que sean “No sabe”.

### 3.3.2. Limpieza general

Tal y como se explicó anteriormente, se crearán dos conjuntos de datos a partir de los cuales se generarán los modelos. Sin embargo, es necesario realizar un paso previo de limpieza, codificación y análisis de las tablas. Cabe destacar que todas las decisiones de limpieza y codificación han sido consensuadas posteriormente con la SES para verificar su coherencia.

En primer lugar, se estudia la información contenida en la tabla *Autor*. Se eliminan 11 casos para los que no existe ningún autor asociado. Se eliminan también otros 6 casos en los que no se conoce la fecha de nacimiento del autor y, por lo tanto, no es posible determinar su edad. A su vez, analizando la tabla *Víctimas*, se eliminan 2 casos en los que se desconoce la fecha de nacimiento de la víctima. Detectamos también otros 440 casos en los que la edad de la víctima registrada y/o la del autor es inferior a 16 años o superior a 90 años. Consideramos que este dato es erróneo y, una vez efectuada toda la limpieza, corregimos dichas edades de víctimas y autores asignándoles respectivamente la mediana de la edad de las víctimas y la mediana de la edad de los autores del resto de casos válidos.

A continuación, analizamos la información contenida en la tabla *Denuncias*. Se eliminan 361 casos en los que no se ha registrado ninguna denuncia, al tratarse de simulaciones de delito. Adicionalmente, se borran 12 casos en los que no se conoce la provincia donde se ha efectuado la denuncia.

Proseguimos con la tabla *Hechos*. Eliminamos 4 casos en los que la localidad registrada era "municipio en pruebas 000". Se entiende que se trata de casos de prueba que se han introducido en la base de datos. En los 486 casos en los que, o bien la localidad registrada no se correspondía con ninguna localidad española, o bien no había ninguna localidad registrada, una vez efectuada toda la limpieza, se ha utilizado como número de habitantes de la localidad la mediana del número habitantes de las localidades donde han ocurrido los hechos del resto de casos válidos.

Al examinar la información contenida en la tabla *Casos*, detectamos 82 casos que se han denunciando en una institución perteneciente a los Mozos de Escuadra. Se eliminan dichos casos, dado que el sistema VioGén es de uso opcional en Cataluña y, por tanto, los casos de esta institución no tienen por qué tener seguimiento. Sin embargo, se observan 64 casos con seguimiento.

A continuación, suprimimos los casos en los que haya algún formulario VPR o VPER para el que no se conozca el nivel de protección asignado o el nivel de protección que el actual sistema VioGén sugirió. Se eliminan 5 casos con sus correspondientes 5 formularios VPR y 76 formularios VPER.

Seguidamente vemos que hay 908 duplicidades con respecto al identificador del caso en los formularios VPR, perteneciendo dichas duplicidades a 844 casos diferentes. Es decir, para cada uno de estos 844 casos hay registrado más de un formulario VPR. De estos, en los 99 casos en los que entre el primer y el segundo formulario VPR asociados al mismo caso han pasado menos de 24 horas, nos quedamos con el formulario que contenga más información, eliminándose el otro. Entendemos que el formulario que tiene más información es aquel que tiene más respuestas contestadas o en el que las repuestas contestadas manifiesten una situación de peligrosidad mayor. En los 745 casos en los que entre el primer y el segundo formulario han pasado más de 24 horas, nos quedamos con el primero, independientemente de la situación. Estos formularios deberían de ser formularios VPER que no se han registrado como tal (nótese que no se transforman a VPER, ya que la información de ambos formularios no es la misma y faltarían datos relevantes). En las 51 ocasiones en las que existen más de dos formularios VPR asociados al mismo caso, los restantes se eliminan siempre.

Por último, observamos que existen casos en los que han transcurrido menos de 15 horas entre dos formularios VPER. Se entiende que esta situación podría deberse a un error a la hora de registrar los formularios. Para subsanar dicho error, si el segundo formulario es de tipo sin reincidencia y en el primero se detecta reincidencia, eliminamos el segundo, salvo en los casos en los que el nivel de protección asignado en el primero fuera extremo. En caso de que el primer formulario fuera de tipo con reincidencia, pero ninguna de las variables pertinentes (ver sección 3.5) refleje reincidencia, nos quedaremos con el formulario que más información contenga. Eliminamos 376 formularios VPER con respecto a la primera situación y 67 formularios VPER con respecto a la segunda.

## 3.4. Características de la base de datos limpia

Una vez efectuada la limpieza disponemos de 44.655 casos con sus correspondientes 44.655 formularios VPR y 252.689 formularios VPER (de estos últimos, 20.864 formularios son de tipo sin reincidencia y 231.825 con reincidencia). En promedio se registran 5,66 formularios VPER por caso.

### 3.4.1. Información general de cada caso.

Como se mencionó anteriormente, más allá de las respuestas dadas en cada uno de los formularios, cada caso presenta una serie de características específicas, como la edad del agresor y de la víctima o el número de habitantes de la localidad donde se producen los hechos. Por simplificar, se presupone que esta información no cambia a lo largo del caso. Consecuentemente, estos datos complementan a las respuestas dadas en los formularios a la hora de realizar cada una de las predicciones. Con respecto a la información general de los casos, obtenemos las siguientes estadísticas:

**Institución:** del conjunto total de casos, 306 han sido denunciados en instituciones de la Policía Forense, 25.646 en las de la Policía Nacional, 1.236 en las de la Policía Local y 17.467 en las de la Guardia Civil.

**Provincia:** hay 38.537 casos que se han producido dentro de la península y 6.118 casos fuera. El promedio y la mediana del número de habitantes de la provincia donde se han producido los casos son 2.124.188 y 1.122.799 habitantes respectivamente.

**Localidad:** el promedio y la mediana del número de habitantes de localidad donde se han producido los hechos son 418.512 y 79.137 respectivamente. Además, hay 6.195 casos en los que la localidad tiene menos de 10.000 habitantes, 13.630 en los que la localidad tiene entre 10.000 y 59.999 habitantes, 7.145 en los que la localidad tiene entre 60.000 y 125.000 habitantes y 17.685 en los que la localidad tiene más de 125.000 habitantes.

**Autor:** el promedio y la mediana de la edad del agresor son respectivamente 38,88 y 38 años. Además, hay 325 agresores que están presentes en más de un caso diferente.

**Víctima:** el promedio y la mediana de la edad de la víctima son respectivamente 36,08 y 35 años. Por otra parte, hay 333 víctimas que están presentes en más de un caso diferente.

En la sección E.1 se estudia la relación entre estos factores y la probabilidad de reincidencia. Adicionalmente, en la sección 3.8 se volverá a analizar su relevancia mediante el cálculo de correlaciones, esta vez con la información ya codificada.

### 3.4.2. Nivel de protección asignado e inactividad de los casos

Cada vez que se rellena un formulario, se valora el riesgo de reincidencia que existe, clasificándose este dentro de uno de los cinco niveles, y se establecen las medidas de protección conforme a ese nivel. El número de formularios VPR y VPER a partir de los cuales se asigna cada uno de los niveles de protección se recoge en la tabla 3.1. Nótese que estos valores son los asignados por la AC y pueden diferir de los predichos por el modelo actual de VioGén. En particular, el modelo VioGén recomienda un nivel de protección inferior al finalmente asignado por la AC en un 6,2 % de los casos y un nivel superior en un 2,2 % de los casos.

| | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **VPR** | 14.122 | 16.484 | 10.483 | 2.949 | 617 |
| **VPER** | 108.527 | 94.033 | 41.208 | 7.519 | 1.402 |

**Tabla 3.1:** Recuento de los niveles de protección aplicados a partir de los formularios.

Por otro lado, los casos pueden declararse inactivos por orden judicial, suprimiéndose las medidas cautelares que se hubiesen establecido. Esto puede ocurrir, por ejemplo, si el agresor entra en prisión o se considera que ya no existe riesgo. No obstante, el caso se puede reactivar posteriormente, si ocurre algún suceso que haga cambiar la situación. Un 89,9 % de los casos han estado inactivos en algún momento. De entre los casos inactivados, el promedio de veces en los que el caso se ha establecido como inactivo es 1,08. Este dato refleja que en la mayoría de casos, después de desactivarse por primera vez, no es necesario reactivarlos en algún momento posterior. El estudio en detalle de estos casos podría ser particularmente interesante, ya que se han archivado y, en consecuencia, se han cancelado las medidas de protección. Dicho estudio se deja para trabajo futuro.

## 3.5. Tipos de reincidencia y delito

A partir de las respuestas de los formularios VPER, no solo podemos detectar si ha habido reincidencia o no, sino que también podemos profundizar en el estudio del tipo de reincidencia que se está dando. A continuación detallamos estos conceptos y los ilustramos con diferentes estadísticas.

De acuerdo con la SES, consideramos que ha habido reincidencia desde la última vez que la

víctima acudió a la AC, si al rellenar el formulario VPER se detecta que ha habido quebrantamientos de órdenes, violencia, amenazas o uso de armas. Adicionalmente, consideraremos los insultos, violencia física y agresiones sexuales como subtipos de reincidencia violenta. Téngase en cuenta que a partir de las respuestas dadas en los formularios VPR, se puede conocer el tipo de delito inicialmente cometido, pero no pueden registrarse quebrantamientos de medidas de protección, puesto que es la primera que la víctima denuncia al agresor. Las respuestas de los formularios VPR y VPER que se utilizan para conocer el tipo de delito inicial y el tipo de reincidencia se presentan en el anexo C.

Diremos que hay reincidencia en el periodo posterior a un formulario VPR o VPER, si en la ventana temporal ligada al nivel de protección asignado hay reincidencia. Las longitudes de las ventanas temporales se corresponden con los plazos máximos hasta la siguiente revisión (ver tabla 2.1). A modo de ejemplo, si el nivel de protección asignado a partir del formulario es bajo, se dice que hay reincidencia en el periodo posterior si hay reincidencia en los próximos 60 días. Si el riesgo asignado es no apreciado, consideramos que hay reincidencia en el periodo posterior, si en algún momento después de rellenarse el formulario hay reincidencia. Esta última definición es importante, dado que nuestro objetivo principal es, cada vez que se rellene un formulario, predecir el nivel de protección con el que evitaríamos que haya reincidencia en la siguiente ventana temporal. Hay que destacar que pueden producirse revisiones antes de que finalice la ventana temporal. A modo de ejemplo, cabe la posibilidad de que se le haya asignado a la víctima riesgo bajo (por lo que el plazo máximo hasta la siguiente revisión es de 60 días) y, sin embargo, se realice una revisión a los 15 días, en la que se cambie el nivel de protección.

Consecuentemente, si al asignar un nivel de protección no hay reincidencia, en la siguiente ventana temporal, será, o bien porque ese nivel es suficiente por sí solo para proteger a la víctima hasta el plazo máximo (e incluso en algunos casos se podría en cierto momento hasta bajar, tal y como se ilustra en la figura 3.1(a)), o bien porque con ese nivel la AC ha sido capaz de detectar un aumento de peligrosidad e incrementar a tiempo el nivel de protección (tal y como se ilustra en la figura 3.1(b)). Por el contrario, si hay reincidencia en la siguiente ventana temporal será, o bien porque el nivel asignado no es suficiente para poder proteger a la víctima (e incluso con alguno de los niveles superiores tampoco se podría proteger, como se ilustra en la figura 3.1(c)) o bien porque con ese nivel de protección no se ha hecho un seguimiento correcto del caso y, a partir de cierto momento, se han reducido las medidas de protección, infravalorando el riesgo (tal y como se ilustra en la figura 3.1(d)).

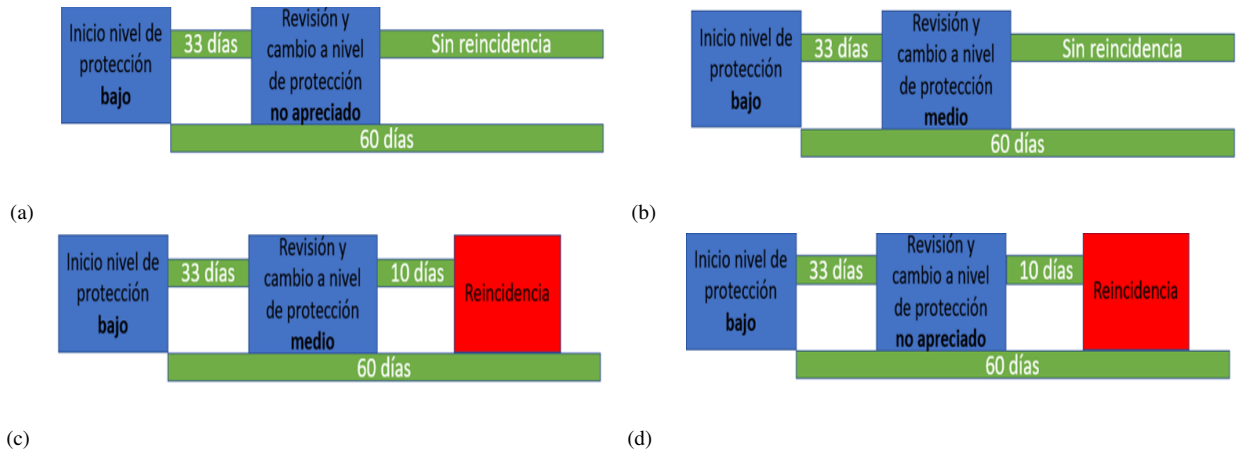


**Figura 3.1:** Ejemplos de situaciones que se pueden dar en la siguiente ventana temporal.

De los 44.655 casos, en 9.086 ha habido algún tipo de reincidencia a lo largo del caso. El promedio de reincidencias para los casos en los que hay al menos una es de 1,67. La probabilidad de que haya reincidencia en el periodo posterior a un formulario VPR se recoge en la tabla 3.2. Se muestra la probabilidad en general y en función del nivel de protección aplicado. Se realiza el desglose de probabilidades para cada subtipo de reincidencia.

| | **Prob general** | **Prob no apreciado** | **Prob bajo** | **Prob medio** | **Prob alto** | **Prob extremo** |
|---|---|---|---|---|---|---|
| **Reinc General** | 0,0855 | 0,1384 | 0,0659 | 0,0593 | 0,0393 | 0,0632 |
| **Reinc sin contar quebrantamientos** | 0,0562 | 0,1105 | 0,0355 | 0,0266 | 0,0190 | 0,0470 |
| **Quebrantamientos** | 0,0467 | 0,0583 | 0,0422 | 0,0442 | 0,0298 | 0,0292 |
| **Reinc violenta** | 0,0451 | 0,0978 | 0,0249 | 0,0170 | 0,0094 | 0,0259 |
| **Reinc con uso de armas** | 0,0053 | 0,0113 | 0,0031 | 0,0020 | 0,0014 | 0,0049 |
| **Reinc con amenazas** | 0,0364 | 0,0647 | 0,0261 | 0,0202 | 0,0156 | 0,0405 |
| **Reinc con insultos** | 0,0364 | 0,0772 | 0,0208 | 0,0143 | 0,0088 | 0,0227 |
| **Reinc con violencia física** | 0,0321 | 0,0748 | 0,0162 | 0,0085 | 0,0037 | 0,0178 |
| **Reinc con agresión sexual** | 0,0021 | 0,0051 | 0,0007 | 0,0008 | 0,0003 | 0,0049 |

**Tabla 3.2:** Probabilidad de reincidencia en el periodo posterior a un formulario VPR en función del nivel de protección asignado. Desglose para cada subtipo de reincidencia.

La probabilidad de que haya reincidencia en el periodo posterior a un formulario VPR, cuando el caso se inactiva dentro de este periodo es de 0,0916.

La probabilidad de que haya reincidencia en el periodo posterior a un formulario VPER se recoge en la tabla 3.3. Se muestra la probabilidad en general y en función del nivel de protección aplicado. Se realiza el desglose de probabilidades para cada subtipo de reincidencia.

| | Prob general | Prob no apreciado | Prob bajo | Prob medio | Prob alto | Prob extremo |
|---|---|---|---|---|---|---|
| **Reinc General** | 0,0762 | 0,0959 | 0,0499 | 0,0814 | 0,0826 | 0,1320 |
| **Reinc sin contar quebrantamientos** | 0,0422 | 0,0602 | 0,0235 | 0,0361 | 0,0411 | 0,0877 |
| **Quebrantamientos** | 0,0551 | 0,0617 | 0,0403 | 0,0681 | 0,0672 | 0,0927 |
| **Reinc violenta** | 0,0325 | 0,0493 | 0,0171 | 0,0243 | 0,0259 | 0,0621 |
| **Reinc con uso de armas** | 0,0038 | 0,0056 | 0,0020 | 0,0032 | 0,0041 | 0,0143 |
| **Reinc con amenazas** | 0,0290 | 0,0407 | 0,0158 | 0,0265 | 0,0327 | 0,0749 |
| **Reinc con insultos** | 0,0272 | 0,0416 | 0,0138 | 0,0196 | 0,0218 | 0,0571 |
| **Reinc con violencia física** | 0,0205 | 0,0317 | 0,0104 | 0,0150 | 0,0125 | 0,0357 |
| **Reinc con agresión sexual** | 0,0015 | 0,0025 | 0,0006 | 0,0007 | 0,0009 | 0,0050 |

**Tabla 3.3:** Probabilidad de reincidencia en el periodo posterior a un formulario VPER en función del nivel de protección asignado. Desglose para cada subtipo de reincidencia.

La probabilidad de que haya reincidencia en el periodo posterior a un formulario VPER, cuando el caso se inactiva dentro de este periodo es de 0,0904. A partir del análisis de estas tablas, podemos ver que, mientras que para los formularios VPR la mayor probabilidad de reincidencia se obtiene cuando se aplica nivel de protección no apreciado, para formularios VPER se obtiene cuando se aplica nivel de protección extremo. Este hecho se puede deber a que en los formularios VPER ya se tiene más información sobre los comportamientos del agresor, de acuerdo al seguimiento que se ha realizado sobre el caso. Por lo tanto, si para un caso se sigue considerando que el nivel de protección que necesita la víctima es extremo, será porque el agresor es excesivamente peligroso. Adicionalmente, en términos generales, la probabilidad de reincidencia es mayor cuando se archiva el caso, que durante una ventana temporal en la que la víctima esta protegida.

## 3.6. Almacenamiento del histórico del caso en formularios VPER

Tal y como se ha explicado anteriormente, de forma novedosa, para cada uno de los formularios VPER registrados vamos a almacenar la evolución del caso hasta ese momento. Para conseguir dicho objetivo, se utilizará la información recogida en cada uno de los formularios VPER anteriores y en el formulario VPR. Cabe destacar que este proceso se realiza una vez la base de datos está limpia.

**Aumentos y decrementos**: En primer lugar, se almacenará la evolución de las respuestas con respecto al formulario VPER anterior. Guardando esta información podemos saber si la situación empeora, mejora o se mantiene estable. Para cada una de las preguntas vamos a guardar si la respuesta

ha aumentado, decrementado o es la misma. Para las respuestas a preguntas de tipo A se considera que la respuesta entre dos formularios ha aumentado si se pasa de “No” a “Sí” y ha decrementado si se pasa de “Sí” a “No”. Para las respuestas a preguntas de tipo B se considera que ha aumentado si se pasa de “No” a “No sabe”, de “No” a “Sí” o de “No sabe” a “Sí” y ha decrementado si se pasa de “Sí” a “No sabe”, de “Sí” a “No” o de “No sabe” a “No”. Los aumentos y decrementos en las respuestas a preguntas de tipo C se consideran de igual manera que las de tipo B, sustituyendo “No sabe” por “No procede”. En las preguntas de tipo E se estudia por separado cada una de las opciones y para cada opción se considera que la respuesta ha aumentado si pasa de no estar seleccionada la opción a estar seleccionada y ha decrementado si pasa de estar seleccionada a no estarlo. En el resto de tipos de preguntas la definición de aumento y decremento está implícita.

Adicionalmente, se registrarán también los aumentos y decrementos con respecto a la reincidencia y cada uno de sus subtipos. Se considera que se ha producido un aumento si se pasa de no detectar reincidencia a detectar y se ha producido un decremento si se pasa de detectar reincidencia a no detectar.

**Recuento de los niveles de protección asignados**: Para cada uno de los formularios VPER se almacenará el número de veces que cada nivel de protección ha sido asignado en ese caso, desde que se registra el formulario VPR inicial hasta el formulario VPER inmediatamente anterior al formulario VPER en cuestión.

De manera complementaria, se guarda para cada formulario VPER el nivel de protección asignado en el formulario VPER inmediatamente anterior y el asignado en el formulario VPR. De esta manera, podemos hacernos una idea de cómo ha variado el riesgo desde que se denuncia el caso hasta el momento en el que se registra el formulario VPER. Guardaremos también el número de formularios VPER que se han registrado previamente en el caso.

La relevancia que tienen las variables, que reflejan la evolución del caso, se muestra en la sección 3.8, una vez se ha codificado la base de datos.

## 3.7. Nivel de protección óptimo e información a predecir

En relación con la reincidencia y el nivel de protección asignado a la víctima se define el nivel de protección óptimo (NPO) . Para cada uno de los formularios que tenemos registrados, si no ha habido reincidencia en la ventana temporal, consideramos que el NPO era el nivel de protección asignado. En caso de que haya habido reincidencia en la ventana temporal, consideramos que el NPO hubiese sido el menor nivel de protección para el cual no hubiese habido reincidencia en la ventana temporal. Nótese que si ha habido reincidencia aplicándose nivel de protección extremo, este es, aún así, el NPO, puesto que no hay ningún nivel superior. A modo de ejemplo, si a la víctima se le asignó nivel de protección bajo y el agresor reincidió a los diez días de rellenarse el formulario, se considera que

el NPO hubiese sido alto. Por el contrario, si a la víctima se le asignó nivel de protección medio y en la ventana temporal no hubo reincidencia, se considera que este era el NPO, dado que no podemos saber si con un nivel de protección inferior hubiese sido suficiente. Por otro lado, en la sección 3.5 hemos definido varios subtipos de reincidencia. Por lo tanto, podemos definir también cuál hubiese sido en cada caso el NPO para evitar uno de los subtipos de reincidencia en concreto. A modo de ejemplo, el NPO para agresión sexual, sería el menor nivel de protección para el que sabemos que, de haberse aplicado, se hubiesen evitado las agresiones sexuales en la siguiente ventana temporal. La idea es dar la máxima información posible a la AC, para que esta pueda decidir el nivel de protección que asigna a la víctima.

| | **No apreciado** | **Bajo** | **Medio** | **Alto** | **Extremo** |
|---|---|---|---|---|---|
| **Reinc General** | 12.167\|0 | 16.808\|2.694 | 10.464\|1.706 | 3.834\|1.188 | 1.382\|782 |
| **Reinc sin contar quebrantamientos** | 12.561\|0 | 17.005\|2.485 | 10.542\|1.533 | 3.433\|759 | 1.114\|517 |
| **Quebrantamientos** | 13.299\|0 | 16.222\|1.812 | 10.385\|1.510 | 3.604\|946 | 1.145\|546 |
| **Reinc violenta** | 12.741\|0 | 17.046\|2.381 | 10.576\|1.463 | 3.332\|632 | 960\|364 |
| **Reinc con uso de armas** | 13.962\|0 | 16.515\|\|1.460 | 10.507\|1.221 | 3.006\|271 | 665\|73 |
| **Reinc con amenazas** | 13.209\|0 | 16.646\|1.983 | 10.487\|1.408 | 3.309\|629 | 1.004\|407 |
| **Reinc con insultos** | 13.031\|0 | 16.883\|2.161 | 10.549\|1.410 | 3.273\|571 | 919\|323 |
| **Reinc con violencia física** | 13.066\|0 | 16.979\|2.159 | 10.588\|1.372 | 3.202\|481 | 820\|225 |
| **Reinc con agresión sexual** | 14.050\|0 | 16.517\|1.402 | 10.482\|1.175 | 2.966\|231 | 640\|49 |

**Tabla 3.4:** Número de veces que se tendría que haber aplicado cada uno de los niveles de protección a partir de los formularios VPR, en función del tipo de reincidencia a evitar.

En la tabla 3.4 se presenta el número de veces en los que se tendría que haber aplicado cada uno de los niveles de protección a partir de los formularios VPR, de acuerdo a la definición del NPO. En rojo se muestran los casos en los que el sistema Viogén recomendó un nivel de protección inferior al NPO. En la tabla 3.5 se presenta el número de veces en los que se tendría que haber aplicado cada uno de los niveles de protección a partir de los formularios VPER, de acuerdo a la definición del NPO. En ambas tablas, se desglosa en función del subtipo de reincidencia que se quiere evitar.

Consecuentemente, el objetivo es, dado un nuevo formulario VPER o VPR, predecir correctamente el NPO. Es decir, queremos predecir con qué nivel podemos evitar que haya reincidencia en la siguiente ventana temporal, intentando consumir a la vez el menor número de recursos posibles. No obstante, dentro de los límites de recursos (ajenos a este estudio) siempre es preferible sobreproteger a la víctima a albergar la posibilidad de que el nivel de protección se quede corto. Nótese que si no hay

reincidencia, salvo que hayamos asignado nivel de protección no apreciado, nunca podremos verificar si el nivel de protección ha sido estrictamente hablando óptimo o, si con un nivel de protección inferior, podríamos haber protegido también a la víctima. Sin embargo, la definición de NPO propuesta para los casos a posteriori, es una aproximación razonable, con la que se optimizan en cierta medida los recursos. De tal manera que, si somos capaces de crear un modelo que prediga correctamente el NPO de acuerdo a nuestra definición para los casos que ya tenemos registrados, si se siguen la recomendaciones del modelo en nuevos casos reales, aumentará la probabilidad de que no haya reincidencia.

Por otro lado, podemos cambiar el enfoque y predecir cual sería la probabilidad de reincidencia en la siguiente ventana temporal, en función del nivel de protección que se aplicase. A partir de estas probabilidades, se podría predecir también el nivel de protección más adecuado para la víctima. Adicionalmente, podemos predecir si habría reincidencia de un subtipo en concreto. Este segundo enfoque puede ser especialmente útil, si queremos tener en cuenta en los modelos el nivel de protección que se asignó a la víctima en los casos registrados. Es fácil entender, en términos de reincidencia, que no es lo mismo que no se produzca reincidencia porque estamos asignando niveles de protección altos, que porque el caso en sí carezca de riesgo independientemente del nivel de protección que se le asigne a la víctima. El uso de cada uno de estos enfoques se verá más detalle en el capítulo 4.

| | **No apreciado** | **Bajo** | **Medio** | **Alto** | **Extremo** |
|---|---|---|---|---|---|
| **Reinc General** | 98.120\|0 | 96.957\|18.763 | 40.969\|9.867 | 11.905\|7.239 | 4.738\|3.400 |
| **Reinc sin contar quebrantamientos** | 101.996\|0 | 96.406\|1.6478 | 41.108\|8.843 | 9.846\|5.066 | 3.333\|2.016 |
| **Quebrantamientos** | 101.826\|0 | 94.658\|16.127 | 40.745\|9.317 | 11.259\|6.528 | 4.201\|2.879 |
| **Reinc violenta** | 103.180\|0 | 96.155\|15.1743 | 41.251\|8.566 | 9.301\|4.467 | 2802\|1.501 |
| **Reinc con uso de armas** | 107.923\|0 | 94.185\|12.617 | 41.152\|7.563 | 7.808\|2.853 | 1.621\|360 |
| **Reinc con amenazas** | 104.107\|0 | 95.468\|15.076 | 40.981\|8.408 | 9.216\|4.411 | 2.917\|1.606 |
| **Reinc con insultos** | 104.010\|0 | 95.854\|15.243 | 41.226\|8.385 | 8.992\|4.139 | 2.607\|1.313 |
| **Reinc con violencia física** | 105.087\|0 | 95.382\|14.463 | 41.233\|8.174 | 8.713\|3.808 | 2.274\|990 |
| **Reinc con agresión sexual** | 108.253\|0 | 94.159\|12.482 | 41.204\|7.457 | 7.605\|2.628 | 1.468\|219 |

**Tabla 3.5:** Porcentajes de cada uno de los niveles de protección óptimo para VPER en función del tipo de reincidencia a evitar.

## 3.8. Análisis de datos codificados

Una vez se ha limpiado la base de datos y generado el histórico para los formularios, VPER se realiza el proceso de codificación de variables. La codificación realizada se presenta en detalle en el anexo D. En esta sección, analizaremos los dos conjuntos de variables codificadas, examinado la importancia de cada variable. Dichos conjuntos serán los utilizados posteriormente para generar los modelos.

En primer lugar, vemos que existen variables de baja activación. Es decir, variables que, salvo en un porcentaje pequeño de casos, toman siempre el mismo valor. Las variables de baja activación son candidatas a ser eliminadas. Estas variables se detallan en la sección E.3 (ver anexo E). Con respecto a las variables asociadas a un formulario VPR, vemos que hay 41 variables que toman el mismo valor en más de un 95 % de los casos. Dentro de este conjunto de variables encontramos: i) variables que indican si la edad de la víctima o del autor esta dentro de un intervalo específico (por ejemplo *edadVictima.76-80*); ii) variables que indican si se ha seleccionado una opción de una pregunta multiopción (por ejemplo, *F02021.IA*); iii) variables que indican si las respuestas a las preguntas del bloque 10 son afirmativas (por ejemplo, *F10010.S*, que indica si la víctima tiene alguna discapacidad). Con respecto a las variables asociadas a un formulario VPER, vemos que hay 214 variables que toman el mismo valor en más de un 95 % de los casos. Observamos que, dentro de este conjunto, se encuentran muchas de las variables que indican si se produce un aumento o un decremento en una respuesta con respecto al formulario anterior. Esto nos indica que no es frecuente que haya mucha variación entre las respuestas de un formulario VPER y el siguiente formulario VPER del caso.

Por otro lado, también se ha calculado la correlación entre cada una de las variables predicativas y el NPO. Las correlaciones se muestran en la sección E.2. Las variables con escasa correlación son también candidatas a ser eliminadas. Con respecto a los formularios VPR, podemos ver como el nivel de gravedad de las amenazas o el nivel de los insultos (*F03001* y *F01011*), tienen una alta correlación con el NPO. También existe una alta correlación con respecto a los celos del agresor (*F05000.S* y *F05000.N*). Adicionalmente, si la víctima piensa que el agresor no es capaz de agredirla con violencia (F12000.N), el NPO disminuye. En contraposición, vemos como el NPO tiene escasa correlación con respecto a indicadores concretos sobre la institución donde se denuncia el caso (*Institucion.PF* e *Institucion.PL*). En cuanto a los formularios VPER, observamos que el NPO tiene una gran correlación, con respecto a si ha habido reincidencia desde la última valoración (*esReinc*), el número de veces que se ha asignado el nivel de protección no apreciado o el nivel de protección medio (*Acum.NA* y *Acum.MD*) y el nivel de protección asignado en la última valoración (*riskProfNum.UltVPER*). Estos resultados nos indican la significancia de incluir la evolución del caso a la hora de realizar las predicciones.

Con respecto a la evolución del nivel de protección que se le asigna a la víctima, en la sección E.4 se recogen una serie de tablas que muestran como de probable es que se pase de un nivel a otro, en función del número de formularios que se han registrado en el caso hasta ese momento. Más

concretamente se muestra como varía el nivel de protección entre el formularios, desde el formulario VPR hasta el cuarto formulario VPER. Se observa como, a medida que pasa el tiempo, aumentan las probabilidades de que la AC vuelva a asignar el mismo nivel de protección que había anteriormente.

# 4

# Modelos de predicción del nivel de protección óptimo

## 4.1. Introducción

Tal y como se explicó en el capítulo anterior, una vez realizada la codificación de la base de datos inicial disponemos de dos conjuntos de datos, el relativo a los formularios VPR y el relativo a los formularios VPER. La Figura 4.1 resume el contenido de las entradas de estos conjuntos de datos.

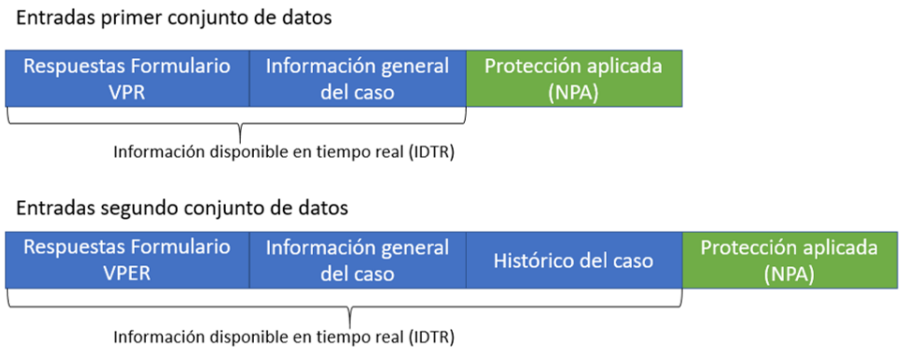


**Figura 4.1:** Información usada para realizar las predicciones

El objetivo es encontrar modelos que, dada la situación actual de un caso y su histórico, predigan cada vez que se rellene un formulario VPR o VPER, el NPO para la víctima. En la sección 3.7 se ha definido cual entendemos que hubiese sido el NPO en cada una de las situaciones registradas. Mediante el primer conjunto de datos predecimos el nivel de protección que hay que asignar cuando se rellene un formulario VPR y, mediante el segundo, cuando se rellene un formulario VPER. Inicialmente contamos con 154 variables para realizar las predicciones a partir de formularios VPR y 292 variables para realizar las predicciones a partir de formularios VPER.

Como en los dos casos las variables a predecir serán las mismas, se probarán los mismos modelos de predicción en ambos, siguiendo el procedimiento presentado en la sección 2.3. Para cada caso y cada modelo en específico, este procedimiento consistirá en: selección de variables con Lasso, balanceo de carga, optimización de hiperparámetros usando validación cruzada y elección del algoritmo que optimiza el modelo.

Cabe destacar que el nivel de protección aplicado en los casos es una variable predicadora especial, puesto que cuando tengamos un nuevo caso real su valor no estará prefijado. Sin embargo, podemos asignarle a la variable un nivel de protección en concreto para predecir que es lo que pasaría en ese caso si protegiésemos a la víctima con ese nivel. Por ejemplo, podríamos predecir si habría reincidencia en caso de que aplicásemos nivel de protección alto.

Para evaluar como de bien predicen los modelos el NPO, se utilizará una métrica específica que penaliza en mayor medida las infravaloraciones que las sobrevaloraciones y tiene en cuenta como de próximo o lejano está el nivel predicho al nivel real. Tal y como se explica en la sección D.4 el NPO se codifica de la siguiente manera: 0 (no apreciado), 1 (bajo), 2 (medio), 3 (alto) y 4 (extremo). Si $Y_1, ...., Y_n$ son los niveles de protección óptimos (NPO's) para cada uno de los casos de nuestro conjunto de validación, el mejor modelo será aquel cuyas predicciones $\hat{Y_1}, ..., \hat{Y_n}$ minimicen el siguiente error:

$$Error(Y_1, .., Y_n, \hat{Y_1}, ..., \hat{Y_n}) = \frac{1}{n}(\sum_{i \in P_1} |Y_i - \hat{Y_i}| + \sum_{i \in P_2} 2|Y_i - \hat{Y_i}|), \tag{4.1}$$

donde $P_1$ es el conjunto de casos en los que el NPO predicho es mayor que el real y $P_2$ el conjunto de casos en los que el NPO predicho es menor que el real. Cabe destacar que los valores predichos serán también discretos. Nótese que hay mayor penalización cuanto mayor es la diferencia entre el riesgo predicho y el real, y además se penaliza el doble las infravaloraciones. Adicionalmente, establecemos inicialmente como condición para los modelos que estos puedan infravalorar como mucho un 12,5 % de los casos. Se incluye esta restricción para intentar reducir el porcentaje de infravaloraciones con respecto al actual sistema Viogén, que infravalora el riesgo un 14,3 % de las veces en formularios VPR y un 15,5 % de las veces en formularios VPER, como se ve en las tablas 3.4 y 3.5. Para los tres primeros modelos (modelos M1, M2 y M3), se elegirán los algoritmos (de entre los presentados en el apartado 2.3.1) e hiperparámetros correspondientes que minimicen el error en la predicción del NPO. En el modelo M4 no utilizaremos esta métrica por razones que se explicarán cuando se presente en detalle el diseño de este modelo en cuestión.

## 4.2. Modelos

Dentro de los modelos de predicción que vamos a utilizar, podemos distinguir entre modelos de predicción basados en la reincidencia y modelos de predicción directos. En los modelos de predicción basados en la reincidencia, lo que predicen los algoritmos es si al aplicar un nivel de protección específico habría reincidencia en la siguiente ventana temporal. A partir de los resultados de reincidencia obtenidos para cada uno de los posibles niveles de protección, se determinaría el NPO para la víctima. Dentro de este enfoque encontramos los modelos M3 y M4, que se explicarán en las secciones 4.2.3 y 4.2.4, respectivamente. En el segundo tipo de modelos los algoritmos predicen directamente el NPO.

Los modelos se entrenan utilizando los algoritmos presentados en el apartado 2.3.1. Cabe destacar que para el segundo tipo de modelos, al no ser la variable respuesta binaria, utilizaremos las versiones multinomiales en los algoritmos de clasificación descritos. Los conjuntos de datos de entrenamiento se balancean de acuerdo a la salida de los algoritmos de predicción. Aunque nos centraremos en predecir el NPO general, es decir, el que pretende evitar cualquier tipo de reincidencia, los modelos se pueden utilizar también para predecir cualquiera de los subtipos de NPO definidos en 3.7.

Los modelos M2, M3 y M4, tendrán en cuenta el nivel de protección asignado a la víctima en la fase de entrenamiento. Nótese que este factor no es tenido en cuenta en los modelos creados por la SES . Estos tres nuevos modelos se han creado con el fin de verificar si la inclusión de este factor, mejora los resultados obtenidos por los modelos en los que no se tiene en cuenta. Podría resultar paradójico incluir esta información en el modelo, cuando lo que queremos predecir es el nivel del protección que hay que asignar a la víctima. En nuevos casos reales podremos utilizar para las predicciones la información disponible en tiempo real (IDTR) , pero no contamos con ninguna información sobre la protección asignada, puesto que esta solo se conoce a posteriori. Sin embargo, el nivel de protección asignado, podría ser un factor clave a la hora de explicar la situación que se ha producido o que podría producirse en caso de aplicarse.

### 4.2.1. Modelo de predicción directa sin incluir la protección: M1

Este modelo trata de predecir el NPO a partir de la IDTR, véase Figura 4.1. Cabe destacar que este modelo es el único que en la fase de entrenamiento no tiene en cuenta el nivel de protección que se le asignó en cada uno de los casos a la víctima. El modelo, cuyo esquema se representa en la figura 4.2, pertenece a la clase de modelos cuyos algoritmos predicen directamente el NPO. El modelo se entrena con el conjunto de datos $\{(IDTR_i, NPO_i)\ ,\ 1 \leq i \leq k\}$, donde $IDTR$ es el conjunto de variables predictoras, $NPO$ la variable respuesta y $k$ el número de entradas. Al introducir una nueva entrada $IDTR_0$ el modelo nos devolverá una predicción del NPO para ese caso, $NPO_0$.

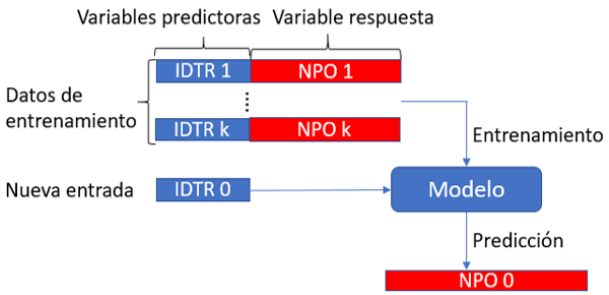


**Figura 4.2:** M1. Modelo de predicción directa sin incluir la protección.

### 4.2.2. Modelo de predicción directa que incluye la protección; M2

Este modelo tiene como entrada la IDTR y el nivel de protección aplicado (NPA) y como salida el NPO. Se trata de una extensión del modelo anterior, en la que sí se tiene en cuenta el NPA. La definición de NPO viene condicionada por el nivel de protección que se le asigno a la víctima. Es decir, si en un caso se aplicó nivel de protección alto y no hubo reincidencia, se considera que ese era el NPO, pero si se hubiese asignado nivel medio y no hubiese habido reincidencia, el NPO hubiese sido medio. En este modelo lo que queremos es predecir NPO|NPA, que se podría interpretar como el NPO que se determinaría a posteriori que hubiese sido el óptimo, si se aplica un nivel de protección en concreto (NPA). Si el NPO fuese superior al NPA significaría que ha habido reincidencia.

Por lo tanto, si para un caso con ciertas características (IDTR), predecimos que aplicando nivel de protección bajo, el NPO sería alto, significaría que este nivel no es suficiente para poder proteger a la víctima. Si por el contrario, se predice que aplicando el nivel de protección bajo, el NPO es bajo o no apreciado, significaría que con este nivel sí se consigue proteger a la víctima. Nótese que si se predice un NPO inferior al NPA lo consideraríamos como si se hubiese predicho el propio NPA, puesto que el NPO siempre tiene que ser igual o superior al NPA. Siguiendo este razonamiento, para cada nuevo caso, del que únicamente conocemos la IDTR, predecimos cual sería el NPO obtenido, si aplicásemos cada uno de los niveles de protección asignables (exceptuando el caso extremo ya que, si se aplica este nivel, siempre va a ser el NPO independientemente de si hay reincidencia).

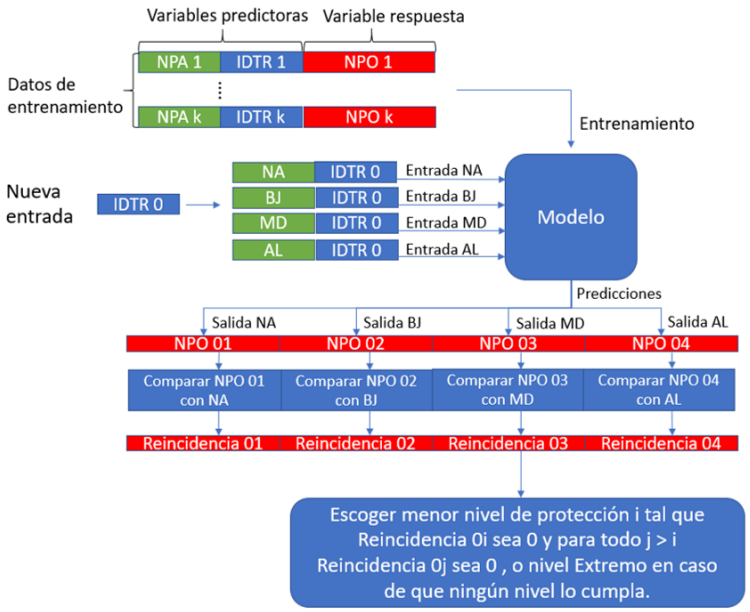


**Figura 4.3:** M2. Modelo de predicción directa que incluye la protección.

Para cada uno de los cuatro supuestos, compararemos el NPA con el NPO que se predice que obtendríamos a posteriori si se aplicase ese nivel. Consideraremos que no habría reincidencia, si el NPA es igual o superior al NPO predicho. Una vez hayamos predicho si habría reincidencia o no,

aplicando cada uno de los cuatro niveles, consideraremos como NPO para el caso el menor nivel para el que se prediga que no habría reincidencia y no haya un nivel de protección superior para el que se prediga reincidencia. En caso de que ningún nivel de protección cumpla estas condiciones, consideraremos que el NPO es extremo. A modo de ejemplo, se predeciría que el NPO es bajo si, para el nivel de protección no apreciado, se considera que va a haber reincidencia y para ninguno de los niveles de protección bajo, medio y alto, se considera que va a haber reincidencia. En la figura 4.3 se puede ver un esquema del modelo.

### 4.2.3. Modelo de predicción basado en la reincidencia que incluye la protección: M3

Este modelo tiene como entrada la IDTR y el NPA y como salida, a diferencia del anterior, si va a haber o no reincidencia en la siguiente ventana temporal. En cada nuevo caso, en el que podemos contar únicamente con la IDTR, predecimos si habría reincidencia de aplicarse cada uno de los cinco posibles niveles de protección. Es decir predecimos a partir del modelo: $Reincidencia|(IDTR, NPA = EX)$,...,$Reincidencia|(IDTR, NPA = NA)$. Para ello creamos cinco entradas añadiendo a la IDTR cada uno de los cinco niveles de protección asignables. Estas entradas serán introducidas en el modelo, el cual nos devolverá la predicción de si habría reincidencia en la siguiente ventana temporal en cada uno de los 5 posibles supuestos.

Una vez hemos predicho si habría reincidencia en caso de que se aplicase cada uno de los cinco posibles niveles, consideramos como NPO el menor nivel para el que se prediga que no hay reincidencia y no tenga un nivel de protección superior para el que se prediga reincidencia. A modo de ejemplo, se predeciría que el NPO es bajo si, para el nivel de protección no apreciado, se considera que va a haber reincidencia y para ninguno de los niveles de protección bajo, medio, alto y extremo, se considera que va a haber reincidencia. En caso de que ningún nivel cumpla tal condición, se considera que el NPO es extremo. En la figura 4.4 se puede ver un esquema del funcionamiento del modelo.

### 4.2.4. Modelo múltiple de predicción basado en la reincidencia que incluye la protección: M4

En este caso, en lugar de incluir el nivel de protección asignado (NPA) como variable predictora en la fase de entrenamiento, se dividen las entradas en función del nivel de protección asignado y se genera un modelo de predicción de reincidencia en la siguiente ventana temporal, para cada uno de los cinco niveles de protección (véase figura 4.5). En cada uno de los modelos se utiliza, para el entrenamiento, el subconjunto de casos cuyo nivel de protección asignado corresponda con el nivel asociado al modelo. Para realizar las predicciones se utilizara únicamente la IDTR. Más concretamente, generaremos un submodelo que nos prediga si hay o no reincidencia cuando se asigna el nivel de

protección no apreciado, otro que nos prediga si hay o no reincidencia cuando se asigna el nivel de protección bajo, y así sucesivamente.

Dada una nueva entrada, se predice si hay reincidencia o no con cada uno de los modelos ligados a cada uno de los cinco niveles de protección. Predecimos como NPO el menor nivel para el que su modelo asociado prediga que no va a haber reincidencia y no haya un nivel de protección superior para el que su modelo prediga que va a haber reincidencia. En cada uno de los submodelos, se utiliza el algoritmo de aprendizaje automático y los hiperparámetros para los que mejor puntuación $F1$ se obtenga a la hora de predecir si ha habido o no reincidencia. Es decir, antes de probar el modelo general, se intenta escoger de forma individual los algoritmos e hiperparámetros que mejor funcionan en cada uno de los submodelos.

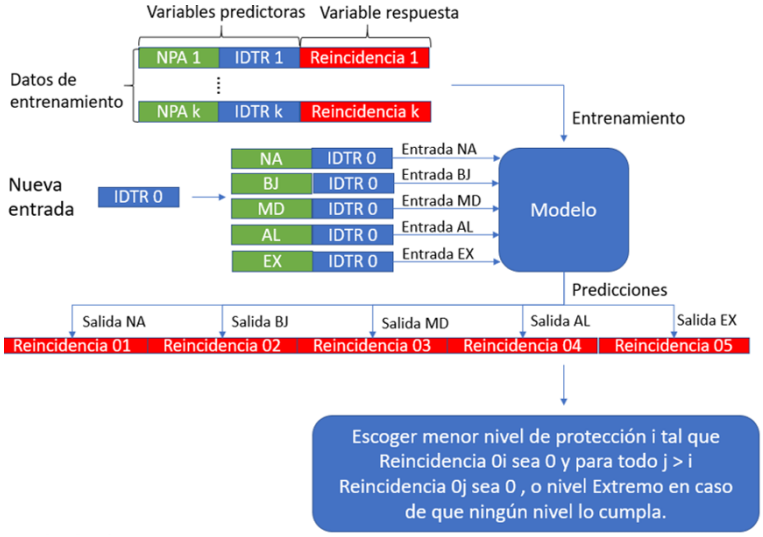


**Figura 4.4:** M3. Modelo de predicción basado en la reincidencia que incluye la protección.

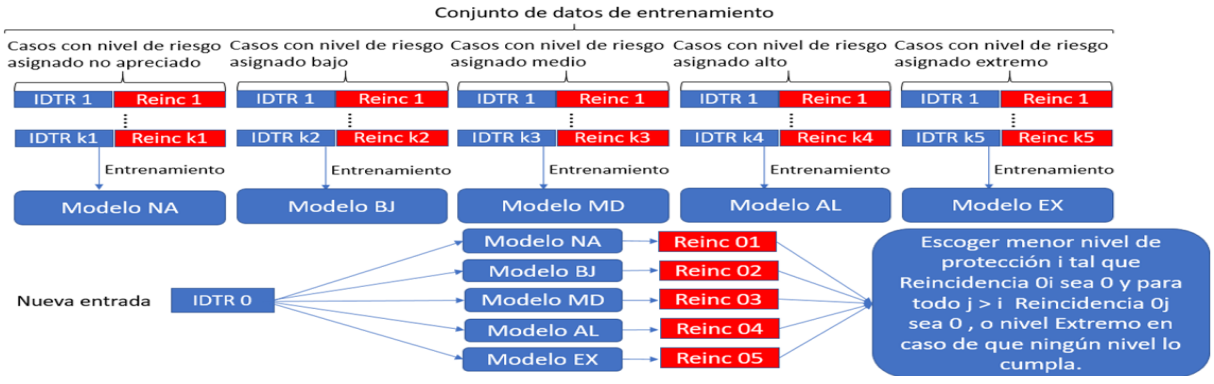


**Figura 4.5:** M4. Modelo múltiple de predicción basado en la reincidencia que incluye la protección.

# 5

# Resultados obtenidos

Tal y como se ha indicado, nuestro principal objetivo era predecir el NPO para evitar cualquier tipo de reincidencia en general. Por tal motivo, nos hemos centrado en validar como funcionan los modelos cuando se predice esta variable. En futuros trabajos se podría comprobar la exactitud de estos modelos a la hora de predecir el NPO con el que se evitase un tipo de reincidencia en concreto.

## 5.1. Resultados M2, M3 y M4: Modelos que tienen en cuenta el nivel de protección aplicado

En términos generales, los modelos que tienen en cuenta el nivel de protección asignado a la víctima en la fase de entrenamiento, modelos M2, M3 y M4, han generado resultados menos satisfactorios. Los mejores resultado obtenidos para estos modelos, se detallan en concreto en el anexo F.

Utilizando estos modelos hemos identificado dos problemas principales. El primero surge al realizar la selección de variables. Tal y como se explicó en el capítulo anterior, para cada uno de los modelos se aplica en primer lugar Lasso, con el objetivo de reducir el número de variables. Lasso es un método de selección de variables que elimina las variables que son menos significativas al modelar el problema con regresión lineal. Sin embargo, en nuestro caso particular, las predicciones obtenidas al modelar el problema de esta manera no son excesivamente buenas. Adicionalmente, analizando más en detalle las variables eliminadas por Lasso en cada modelo (recogidas en E.5), vemos como dentro de los conjuntos de variables eliminadas se encuentran variables que, en teoría, sí deberían de ser importantes a la hora de predecir el riesgo. A modo de ejemplo, en el modelo M3, Lasso elimina del conjunto de variables asociada a los VPR la variable *F01011* (variable que refleja el nivel de gravedad de las agresiones físicas) y, del conjunto de variables asociadas a los VPER, la variable *riskProf.Num.UltVPER* (variable que indica el nivel de protección que se asignó en el último formulario).

En contraposición, al modelar M1 con regresión lineal, sí obtenemos resultados satisfactorios, como veremos en la sección 5.2. Además, se ha podido verificar para este modelo, que el conjunto de variables eliminadas por Lasso es coherente, siendo la mayoría de variables eliminadas, variables con

poca activación o variables con poca correlación con el NPO. Teniendo todos estos factores en cuenta, se ha decidido suprimir para los modelos M2, M3 y M4, las variables eliminadas al aplicar Lasso al modelo M1.

La principal característica que penaliza a estos tres modelos es que para predecir el NPO es necesario realizar cinco predicciones en lugar de una como en el modelo M1. En consecuencia, las predicciones deberían de ser extremadamente precisas para que los modelos funcionasen bien. No obstante, existen otros factores secundarios que también empeoran las predicciones. La tabla 3.1 ilustra el número de veces que se ha aplicado cada nivel de protección. El nivel de protección extremo solo ha sido asignado 617 veces a partir de formularios VPR y 1.402 a partir de formularios VPER. Cabe recordar que para el modelo M4, creamos un submodelo que prediga si habría reincidencia, en caso de que se aplicase nivel de protección extremo. Para obtener tal submodelo, se utiliza en la fase de entrenamiento el conjunto de casos que se han tratado con este nivel de protección. A partir de un conjunto de casos de entrenamiento tan limitado, se antoja muy complicado poder generar el submodelo de forma correcta.

En los modelos M2 y M3 nos encontramos por norma general dos situaciones. En la primera situación, los algoritmos devuelven la misma predicción con independencia del nivel de protección asignado. Esto ocurre, por ejemplo, si los algoritmos asignan muy poco peso en las predicciones al NPA. Por el contrario, en la segunda situación, aunque los algoritmos sí realizan distintas predicciones en función del nivel de protección que se aplique, no modelan correctamente el problema.

Por otro lado, al analizar estos modelos observamos un hecho muy significativo. Si restringimos el estudio de variables a los casos en los que se asigna un nivel de protección en concreto, nos encontramos con que aparecen variables que dentro de ese conjunto son de baja activación. Por ejemplo, la variable *F09030.N* toma el valor 0 en un 99 % de los casos que se tratan como extremos. Es decir, en prácticamente la totalidad de los casos que se tratan como extremos, no se ha podido determinar que el agresor no padezca problemas de adicción. Al estudiar la relación entre este factor y la reincidencia, observamos que la probabilidad de que el agresor reincida, es de 0,277 en los casos en los que no se niega que el agresor padezca problemas de adicción, mientras que esta probabilidad se reduce a 0,162 para casos en los que sí se niega. Por lo tanto, parece que este factor debería de ser importante a la hora de predecir si va a haber reincidencia. Sin embargo, al ser una variable de escasa activación para casos extremos, si entrenamos el submodelo mencionado anteriormente únicamente con estos casos, la mayoría de algoritmos no tendrán en cuenta esta variable. Consecuentemente, cuando queramos predecir en un nuevo caso qué pasaría si se aplicase nivel de protección extremo, este factor no será tenido en cuenta, aún pudiendo ser importante. Del mismo modo, en los casos en los que se ha aplicado nivel de protección no apreciado, la gravedad de los delitos suele ser siempre nula o baja, por lo que nos encontraríamos con un problema similar, al tratar por separado los casos en los que se ha asignado este nivel de protección.

Una de las posibles soluciones que se podría plantear para solucionar este problema, sería utilizar para los modelos que tienen en cuenta el nivel del protección aplicado, una base de datos “artificial”, añadiendo nuevas entradas a la base de datos original. Es decir, si tenemos un caso en el aplicando nivel de protección bajo no ha habido reincidencia en la siguiente ventana temporal, sabemos que si en ese caso se hubiese aplicado un nivel de protección superior tampoco hubiese habido reincidencia. Por lo tanto, para los niveles medio, alto y extremo, podríamos añadir una nueva entrada en la que mantuviésemos todas las características, salvo el nivel de protección aplicado. De la misma manera, si hay reincidencia en un caso en el que se ha aplicado nivel de protección alto, entonces sabemos que también hubiese habido reincidencia si se hubiese aplicado un nivel de protección inferior. De forma análoga al caso anterior, se podría añadir nuevas entradas manteniendo las características del caso, pero simulando que se ha aplicado otro nivel de protección.

## 5.2. Resultados M1: Modelo de predicción directa sin incluir la protección

### 5.2.1. Resultados obtendios al estudiar la reincidencia en general

Dentro del diseño M1, el algoritmo que mejor ha funcionado ha sido la regresión Lasso (con $\lambda = 0,004$), tanto para la predicción del NPO a partir de formularios VPR, como para la predicción a partir de formularios VPER.

Al utilizar regresión Lasso ha sido necesario establecer cortes para determinar a qué nivel de protección corresponde cada predicción. De forma general, la gran mayoría de las predicciones estaban en el intervalo [0,4], por lo que se dividió este intervalo en 5 subintervalos, cada uno de tamaño 0,8 y asociado un nivel. Posteriormente, analizando los resultados, se observó que había demasiados casos en los que se predecía que el NPO era alto y extremo. Consecuentemente, se probó si era posible reducir al error moviendo los dos cortes para acceder a estos niveles de protección. Como las variaciones entre los cortes óptimos del modelo de predicción a partir de formularios VPR y del modelo de predicción a partir de formularios VPER han sido muy pequeñas, se ha decidido establecer el mismo conjunto de cortes en los dos casos. Consecuentemente, se considera que el NPO es no apreciado si la predicción es menor que 0,80, bajo si está en el intervalo [0,80 , 1,60), medio si está en el intervalo [1,60 , 2,60), alto si está en el intervalo [2,60 ,3,48) y extremo si es mayor de 3,48.

En la tabla 5.1 se muestra la matriz de confusión obtenida para el caso de predicción del NPO a partir de los formularios VPR. Los casos por debajo de la diagonal, casos en rojo, son casos en los que se hubiese recomendado un nivel de protección inferior al NPO. De entre los casos representados en verde, podemos distinguir entre: i) casos en los que se hubiese recomendado exactamente el NPO (casos situados en la diagonal); ii) casos en los que se hubiese sobreprotegido a la víctima (casos

situados por encima de la diagonal). El error obtenido de acuerdo a 4.1 es 0,698.

| NPO real \NPO predicho | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | 6122 | 5239 | 790 | 16 | 0 |
| **Bajo** | 1004 | 7595 | 7651 | 553 | 5 |
| **Medio** | 134 | 676 | 6482 | 3114 | 58 |
| **Alto** | 80 | 301 | 1135 | 2054 | 264 |
| **Extremo** | 57 | 198 | 371 | 540 | 216 |

**Tabla 5.1:** Mejores resultados de predicción del riesgo óptimo a partir de formularios VPR

Podemos observar que los resultados son bastante satisfactorios. En la tabla 5.1 se puede ver como nuestro modelo únicamente infravalora el riesgo y asigna un nivel de protección inferior al NPO en un 10,1 % de los casos. Adicionalmente, nuestro modelo solo infravalora con más de un nivel de protección de diferencia con respecto al NPO en un 2,6 % de los casos. En comparación con estos resultados, el sistema VioGén vigente infravalora un 14,3 % de los casos. Además, nuestro modelo hubiese corregido un 31,7 % de los casos en los que el sistema VioGén infravaloró el riesgo y se produjo reincidencia en la siguiente ventana temporal. Para obtener esta estadística, se calcula en cuántos de los casos registrados en los que el sistema Viogen infravaloró el riesgo y hubo reincidencia, nuestro modelo infravalora también el riesgo.

Cabe recordar que para los casos registrados, el NPO es el menor nivel con el que sabemos seguro que no hubiese habido reincidencia (salvo si se aplicó nivel de protección extremo y hubo reincidencia). No obstante, podría ocurrir que con un nivel de protección inferior tampoco hubiese habido reincidencia. Es decir, si se ha asignado nivel de protección alto y no ha habido reincidencia, se considera que el NPO es alto, porque es el nivel de protección para el que se tiene evidencias de que no hay reincidencia. Sin embargo, para un nivel de protección inferior podría darse el caso de que tampoco hubiese habido reincidencia. Por lo tanto, en el 10,1 % de los casos que infravaloramos, no tendría por qué haber reincidencia, sino que a lo sumo, dejaríamos desprotegidas a las víctimas en un 10,1 % de los casos. Por otro lado, se predice el NPO de manera exacta en un 50,3 % de los casos. Adicionalmente, solo sobrevaloramos con más de un nivel de protección de diferencia con respecto al NPO en un 3,1 % de los casos.

Con respecto a la predicción del NPO para formularios VPER, la matriz de confusión obtenida se muestra en la tabla 5.2. El error obtenido de acuerdo a 4.1 es 0,548.

Los resultados de predicción para formularios VPER son incluso mejores que los obtenidos para formularios VPR. Podemos ver en la tabla 5.2 como el modelo solo infravalora un 5,7 % de los casos y solo infravalora con más de un nivel de diferencia con respecto al NPO en un 1,6 % de los casos. En contraposición con estos resultados, el sistema VioGén vigente infravalora un 15,5 % de los casos. Además, nuestro modelo hubiese corregido un 38,7 % de los casos en los que el sistema VioGén infra-

| NPO real \NPO predicho | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | 51579 | 41478 | 4941 | 111 | 11 |
| **Bajo** | 3105 | 59811 | 32610 | 1367 | 64 |
| **Medio** | 428 | 2554 | 28557 | 8733 | 697 |
| **Alto** | 331 | 1336 | 3218 | 5544 | 1476 |
| **Extremo** | 129 | 587 | 1253 | 1561 | 1208 |

**Tabla 5.2:** Mejores resultados de predicción del riesgo óptimo a partir de formularios VPER

valoró el riesgo y se produjo reincidencia en el periodo posterior. Por otra parte, nuestro modelo predice el NPO de manera exacta en un 58,1 % de los casos. Adicionalmente, el modelo solo sobrevalora con más de un nivel de diferencia con respecto al NPO en un 2,8 % de los casos.

### 5.2.2. Resultados obtenidos al estudiar un tipo de reincidencia en concreto

Como comentamos al comienzo del capítulo, se puede predecir también el NPO para evitar un tipo de reincidencia en concreto. De este modo podemos dar a la AC una información más detallada de lo que se predice que puede ocurrir. En especial, nos interesaría calcular el NPO cuando no se consideran los quebrantamientos como reincidencia. Este caso es particularmente interesante ya que los quebrantamientos de órdenes judiciales son el tipo de reincidencia más atenuado, puesto que la víctima no sufre daños directos. Para ver cómo de exactas serían las predicciones en este caso utilizamos, el modelo que nos ha dado mejores resultados al estudiar la reincidencia en general. En la tabla 5.3 se presenta la matriz de confusión obtenida para el caso de predicción del NPO a partir de formularios VPR, cuando no se tienen en cuenta los quebrantamientos. El error obtenido de acuerdo a 4.1 es 0,625. Nótese que en este caso la reincidencia estudiada es más grave que en el caso general.

| NPO real \NPO predicho | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | 7465 | 4313 | 764 | 19 | 0 |
| **Bajo** | 1343 | 8360 | 6749 | 546 | 7 |
| **Medio** | 168 | 678 | 6759 | 2875 | 62 |
| **Alto** | 64 | 193 | 940 | 1984 | 252 |
| **Extremo** | 54 | 134 | 241 | 464 | 221 |

**Tabla 5.3:** Mejores resultados de predicción del riesgo óptimo a partir de formularios VPR, cuando no se tienen en cuenta los quebrantamientos.

El modelo únicamente infravalora el riesgo y asigna un nivel de protección inferior al NPO en un 9,6 % de los casos. En comparación con estos resultados, el sistema VioGén vigente infravalora un 11,9 % de los casos. Además nuestro modelo hubiese corregido un 29 % de los casos en los que el

sistema VioGén infravaloró el riesgo y se produjo reincidencia, sin contar los quebrantamientos, en el periodo posterior. Por otro lado, se predice el NPO de manera exacta en un 55,5 % de los casos.

Con respecto a la predicción del NPO para formularios VPER cuando no se tienen en cuenta los quebrantamientos, la matriz de confusión obtenida se muestra en la tabla 5.4. El error obtenido de acuerdo a 4.1 es 0,453.

| NPO real \NPO predicho | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | 67307 | 31330 | 3251 | 95 | 13 |
| **Bajo** | 3724 | 68394 | 23203 | 1028 | 57 |
| **Medio** | 321 | 2927 | 29504 | 7717 | 639 |
| **Alto** | 227 | 803 | 2276 | 5172 | 1368 |
| **Extremo** | 93 | 355 | 705 | 1073 | 1107 |

**Tabla 5.4:** Mejores resultados de predicción del riesgo óptimo a partir de formularios VPER, cuando no se tienen en cuenta los quebrantamientos.

En este caso, el modelo únicamente infravalora el riesgo y asigna un nivel de protección inferior al NPO en un 4,9 % de los casos. Por contra, el sistema VioGén vigente infravalora un 12,8 % de los casos. Además, nuestro modelo hubiese corregido un 35 % de los casos en los que el sistema VioGén infravaloró el riesgo y se produjo reincidencia, sin contar los quebrantamientos, en el periodo posterior. Por otro lado, se predice el NPO de manera exacta en un 67,8 % de los casos. Podemos concluir que tanto para formularios VPR, como VPER, cuando no se tienen en cuenta los quebrantamientos se infravaloran menos casos y el porcentaje de aciertos es notablemente superior.

# 6

# Conclusiones y líneas futuras

A lo largo de este trabajo se han realizado múltiples avances en la búsqueda de un sistema de predicción alternativo al actual sistema VioGén:

- Se ha realizado el proceso de limpieza de la base de datos inicial, estudiando y corrigiendo las incoherencias entre la información registrada. Dicho proceso ha sido laborioso, dada la envergadura de la base de datos tratada. No obstante, se ha automatizado la mayor parte de este proceso, lo que supone una gran ventaja, puesto que podrá ser reutilizado en futuras extracciones.
- Se han estudiado nuevas variables exógenas con respecto al entorno donde se producen los hechos, como por ejemplo el número de habitantes de la localidad. Podemos destacar que, aunque estas variables no son en principio muy significativas, algunas de ellas sí son tenidas en cuenta por los modelos a la hora de realizar las predicciones (véase sección E.5).
- Se ha almacenado la evolución de los casos hasta el momento anterior a cada formulario VPER. Se ha podido comprobar a la vista de las correlaciones calculadas y de las variables eliminadas por los modelos, que esta información es relevante a la hora de realizar las predicciones.
- Se han creado dos versiones de codificación de la base de datos. Cabe destacar que la codificación de variables también se ha automatizado, por lo que podrá ser reutilizada.
- Se han introducido técnicas de Aprendizaje Automático a la hora de realizar las predicciones.
- Se han diseñado nuevos modelos de predicción y se ha contrastado la eficacia de incluir la protección asignada en la fase de entrenamiento (factor que no se había tenido en cuenta antes), comprobándose con los resultados que, en principio, no se pueden obtener mejoras al incluir este factor.

Como resultado de todo este proceso hemos obtenido un nuevo modelo de predicción que reduce significativamente el número de infravaloraciones con respecto al actual sistema Viogén, tanto para las

predicciones realizadas a partir de los formularios VPR, como para las predicciones realizadas a partir de los formularios VPER.

A la vista de los resultados obtenidos y partiendo del trabajo ya realizado, se proponen varías vías de exploración futuras. Se puede seguir buscando variables exógenas que reflejen el ambiente en el que se produce el caso, como podría ser el porcentaje de parados o la tasa de criminalidad de la localidad donde se producen los hechos. Por otra parte, los resultados obtenidos a la hora de realizar las predicciones a partir de los formularios VPER, nos hacen ver la importancia de reflejar la evolución que ha tenido el caso hasta el momento anterior a cada formulario. Una posible vía de trabajo sería generar información más detallada sobre la evolución de los casos. Se podría hacer también un estudio más extenso sobre la significancia de cada variable con respecto a la reincidencia.

Adicionalmente, de las dos versiones de codificación creadas, durante este trabajo solo hemos utilizado la versión principal. Se podrían validar los modelos con la versión de codificación alternativa, por si hubiese alguna mejora en los resultados.

Con respecto a los modelos, una vez tenemos claro que el modelo que mejor funciona es el que predice el NPO de manera directa, sin tener en cuenta el nivel de protección asignado en la fase de entrenamiento (modelo M1), podemos intentar optimizar los resultados usando métodos *ensemble*, como los presentados en la sección 2.3, para combinar los resultados obtenidos por diferentes algoritmos. Para el resto de modelos que sí tienen en cuenta el nivel de protección asignado, se ha propuesto en la sección 5.1, la creación de una base de datos “artificial”. Se podrían probar los modelos M2, M3 y M4 con esta nueva base de datos, aunque no se prevé que puedan llegar a mejorarse las predicciones obtenidas con M1.

Por último, se podría crear un nuevo modelo que nos prediga la probabilidad de reincidencia cuando se desactiva un caso. De esta manera, además de proporcionar una herramienta que sugiera a la AC el nivel de protección adecuado para la víctima en cada momento, se proporcionaría una herramienta complementaria que indique si no hay riesgo por archivar el caso o, si por el contrario, existe la posibilidad de que haya reincidencia posteriormente.

# Bibliografía


[1] "Violence against women: a global health problem of epidemic proportions, world health organization." http://www.who.int/mediacentre/news/releases/2013/violence_against_women_20130620/en/.

[2] "Ley orgánica 1/2004, del 28 de diciembre, de medidas de protección integral contra la violencia de género."

[3] J. J. L. Ossorio, *Construcción y validación de los formularios de valoración policial del riesgo de reincidencia y violencia grave contra la pareja (VPR4.0 – VPER4.0) del Ministerio del Interior de España*. PhD thesis, Tesis doctoral. Universidad Autónoma de Madrid, 2017.

[4] S. Kampakis, *Predictive modeling of football injuries*. PhD thesis, Department of Computer Science, University College London, 2016.

[5] S. F. Sabbeh, "Machine-learning techniques for customer retention: A comparative study," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 2, 2018.

[6] J. Bachner, "Predictive policing: Preventing crime with data and analytics," tech. rep., Johns Hopkins University, 2013.

[7] L. Cao, "Data science: a comprehensive overview," *ACM Computing Surveys (CSUR)*, vol. 50, no. 3, pp. 1–42, 2017.

[8] "Ministerio de Sanidad, Consumo y Bienestar Social." https://www.mscbs.gob.es/va/ssi/violenciaGenero/Sensibilizacion/AplicacionLibres/home.htm.

[9] "Espacio Digital Ygualex." http://www.ygualex.com/.

[10] "Ministerio del Interior." https://alertcops.ses.mir.es/mialertcops/.

[11] "Fundación Cermi Mujeres." http://www.fundacioncermimujeres.es/es/noticias/ya-puedes-descargarte-la-app-pormi.

[12] "Play google." https://play.google.com/store/apps/details?id=com.google.android.apps. emergencyassist&hl=es.

[13] "Play Google." https://play.google.com/store/apps/details?id=es.seguras&hl=es.

[14] G.Shobha and S. Rangaswamy, "Machine learning," in *Handbook of Statistics* (Elsevier, ed.), vol. 38, ch. 8, pp. 197–228, V.N. Gudivada and C.R. Rao, 2018.

[15] V. Fernández and R. S.-M. Fernández, "Regresión logística multinomial," tech. rep., Departamento de Estadística e Investigación Operativa. Universidad de Valladolid, 2004.

[16] J. W. Grzymala-Busse, "Selected algorithms of machine learning from examples," tech. rep., Department of Computer Science, University of Kansas, 1993.

[17] Z. Zhi-Hua, *Ensemble Methods: Foundations and Algorithms*. Chapman Hall/CRC, 1st ed., 2012.

[18] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization path for generalized linear models by coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, 2010.

[19] P. Peres-Neto, D. Jackson, and K. Somers, “How many principal components? stopping rules for determining the number of non-trivial axes revisited,” *Computational Statistics and Data Analysis*, vol. 49, no. 4, pp. 974–997, 2005.

[20] M. Tenenhaus, V. Vinzi, Y. Chatelin, and C.Lauro, “Pls path modeling,” *Computational Statistics and Data Analysis*, vol. 48, no. 1, pp. 159–205, 2005.

[21] V. V. Asch, “Macro- and micro-averaged evaluation measure,” 2013.

[22] “Instituto nacional de estadística.” «http://www.ine.es/».

# Acrónimos

**AC** autoridad competente.

**IDTR** información disponible en tiempo real.

**NPA** nivel de protección aplicado.

**NPO** nivel de protección óptimo.

**SES** Secretaría de Estado de Seguridad del Ministerio del Interior.

**VPER** formulario de Valoración Policial de la Evolución del Riesgo.

**VPR** formulario de Valoración Policial del Riesgo.

# Apéndices

# A

# Formularios de valoración del riesgo.

## A.1. Formulario de valoración policial del riesgo (VPR)

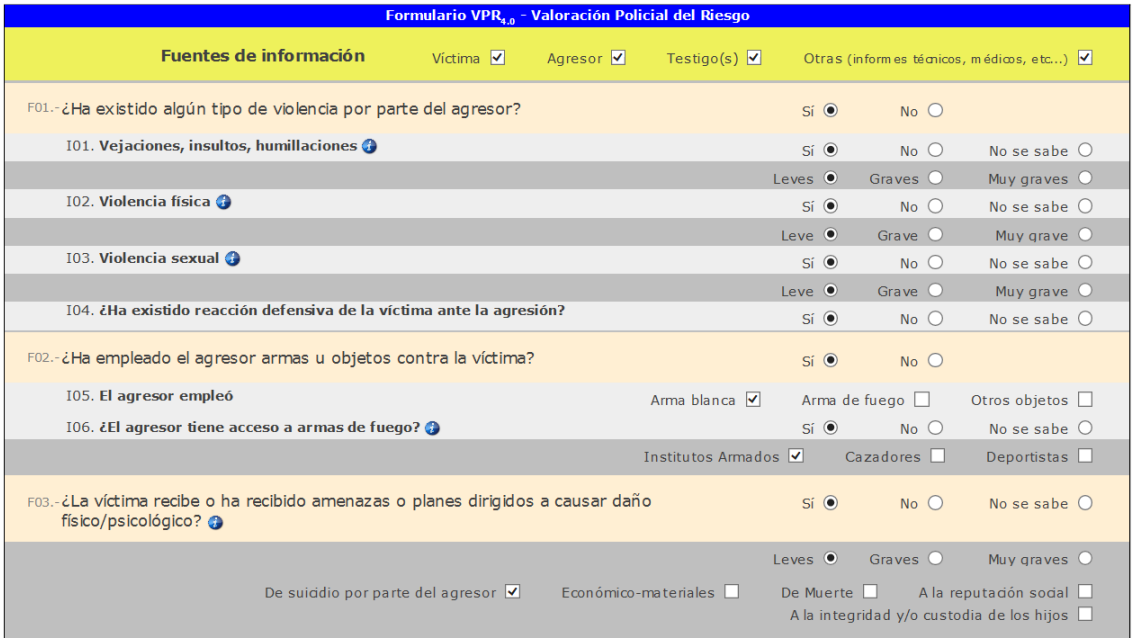


**Figura A.1:** Formulario VPR bloques 1 a 3


**Figura A.2:** Formulario VPR bloques 4 a 8

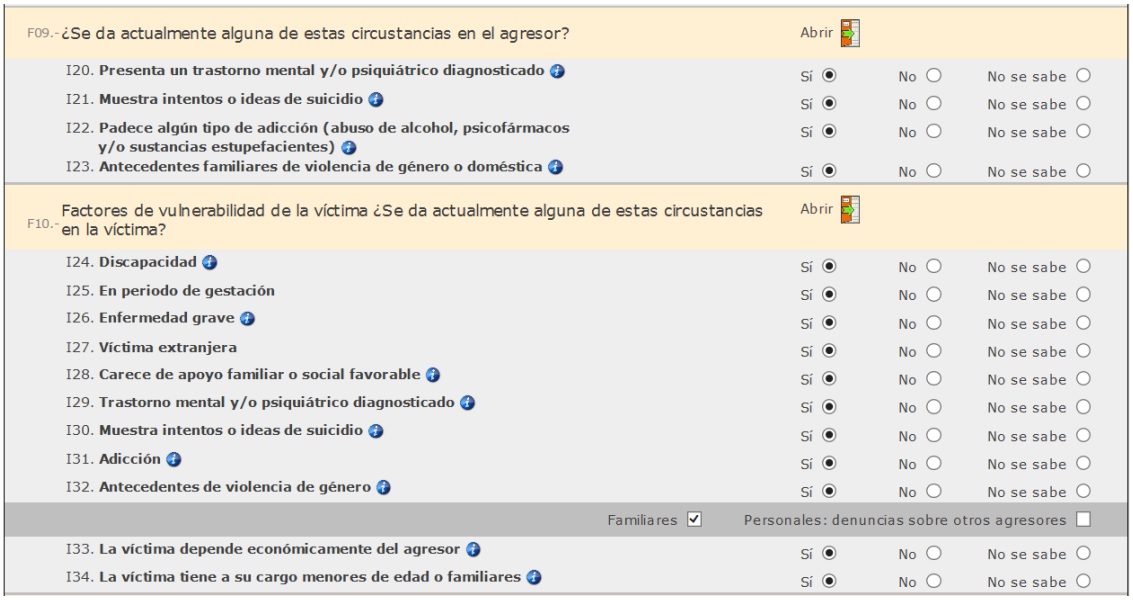


**Figura A.3:** Formulario VPR bloques 9 a 10

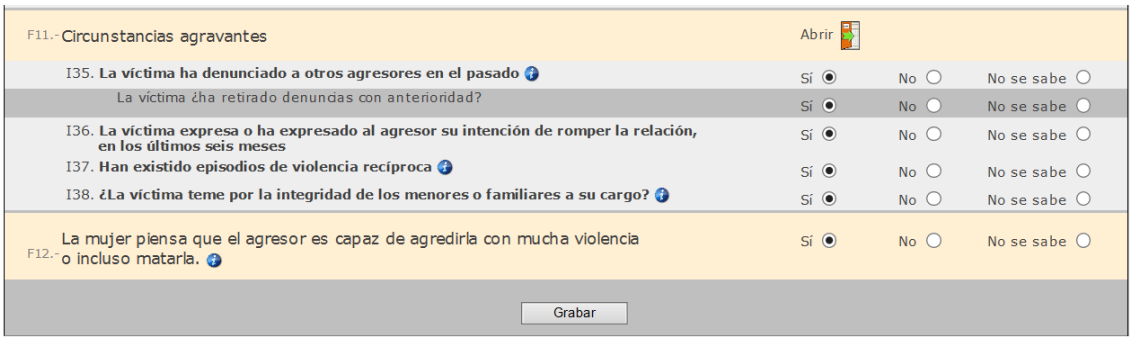


**Figura A.4:** Formulario VPR bloques 11 a 12

## A.2. Formulario de valoración policial del riesgo (VPER)

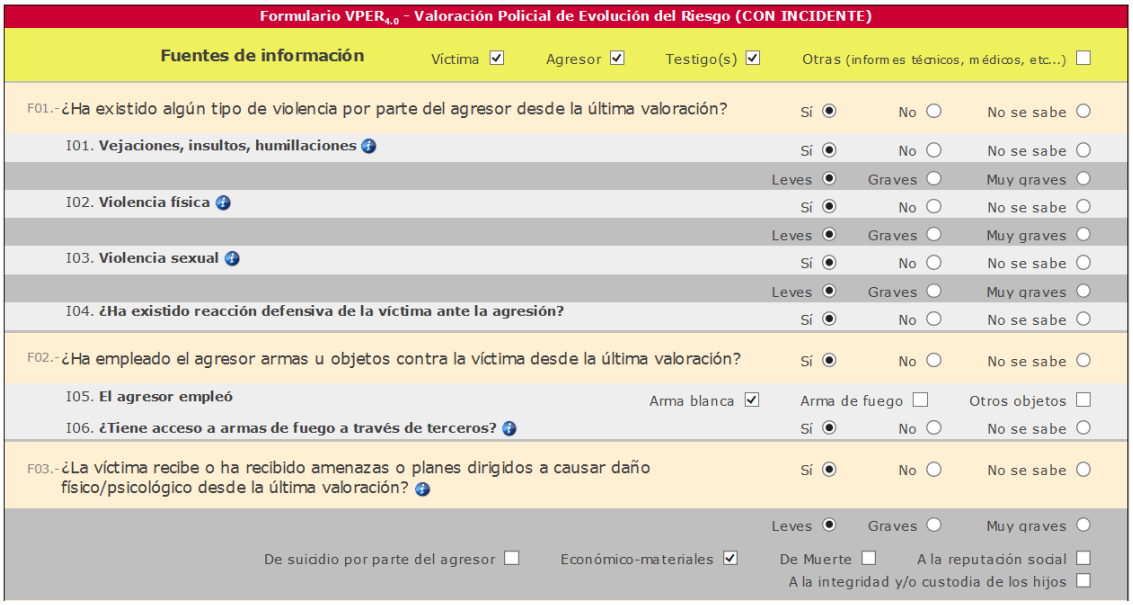


**Figura A.5:** Formulario VPER bloques 1 a 3

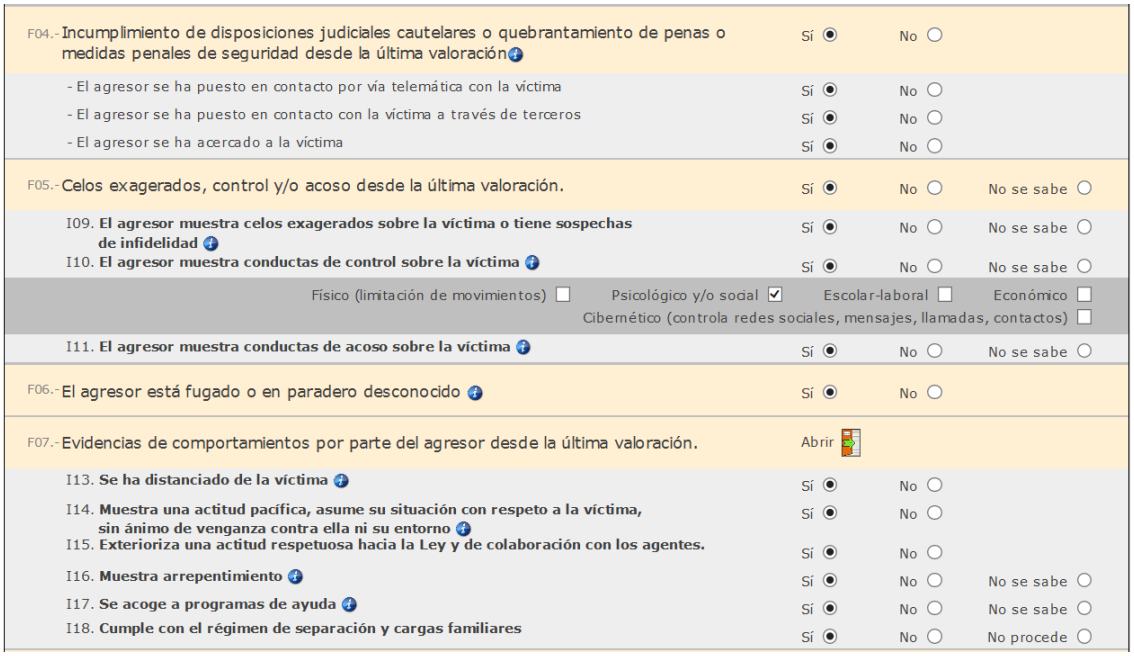


**Figura A.6:** Formulario VPER bloques 4 a 7


**Figura A.7:** Formulario VPER bloques 8 a 11

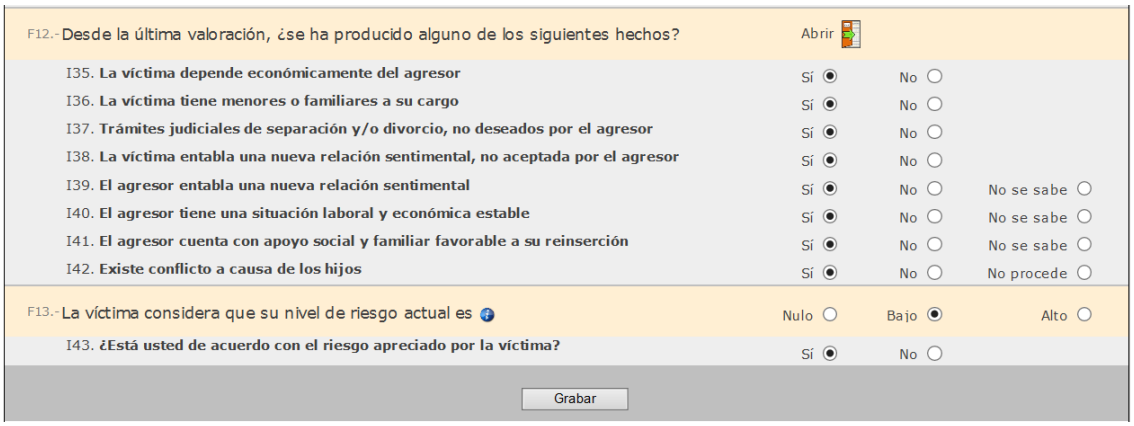


**Figura A.8:** Formulario VPER bloques 12 a 13

# B

# Información contenida en las tablas

En este apartado se detalla el contenido de cada una de las 8 tablas facilitadas por la Secretaría de Estado de Seguridad del Ministerio del Interior. La información y los campos contenidos en cada una de las tablas es la siguiente:

**Autores:** Esta tabla recoge los datos sobres los autores de cada caso. Los campos contenidos son los siguientes:

- **idCaso:** Número identificativo del caso.
- **idAutor:** Número identificativo del autor.
- **fecNacimiento:** Fecha de nacimiento del autor.
- **nacionalidad:** Nacionalidad del autor.
- **profesion:** Profesión del autor.
- **educacion:** Nivel de estudios del autor.
- **sitLaboral:** Situación laboral del autor: desempleado, jubilado, etc.

**Denuncias:** Esta tabla recoge los datos sobre las denuncias realizadas. Los campos contenidos son los siguientes:

- **idCaso:** Número identificativo del caso.
- **idVictima:** Número identificativo de la víctima.
- **idAutor:** Número identificativo del autor.
- **idDenuncia:** Número identificativo de la denuncia
- **fecha:** Fecha en la que se realiza la denuncia.
- **diligencia:** Código interno que asignan las instituciones policiales para gestionar los tramites.
- **unidad:** Descripción de la unidad policial donde se ha puesto la denuncia.
- **provincia:** Provincia donde se ha puesto la denuncia.
- **codProvincia:** Codigo identificativo de la provincia donde se ha puesto la denuncia.

**Víctimas:** Esta tabla recoge los datos sobres los víctimas de cada caso. Los campos contenidos son los siguientes:

- **idCaso:** Número identificativo del caso.
- **idVictima:** Número identificativo de la víctima.
- **fecNacimiento:** Fecha de nacimiento de la víctima.
- **nacionalidad:** Nacionalidad de la víctima.
- **profesion:** Profesión de la víctima.
- **educacion:** Nivel de estudios de la víctima.
- **sitLaboral:** Situación laboral de la víctima: desempleada, jubilada, etc.

**Casos:** Este tabla contiene aspectos generales del caso. Los campos contenidos son los siguientes:

- **idCaso:** Número identificativo del caso.
- **idVictima:** Número identificativo de la víctima.
- **idAutor:** Número identificativo del autor.
- **institucion:** Institución donde se ha dado de alta el caso: Policía Nacional (PN), Policía Local (PL), Policía Foral (PF), Guardia Civil (GC), etc.
- **fecAlta:** Fecha en la que se dio de alta el caso.
- **situacion:** Estado acual del caso: activo o inactivo.

**Hechos:** Esta tabla contiene información sobre los hechos que se han producido. Los campos contenidos son los siguientes:

- **idCaso:** Número identificativo del caso.
- **idVictima:** Número identificativo de la víctima.
- **idAutor:** Número identificativo del autor.
- **idDenuncia:** Número identificativo de la denuncia.
- **idHecho:** Número identificativo de los hechos.
- **fecha:** Fecha en la que ocurrieron los hechos.
- **tipo:** Tipo de hechos que se produjeron: amenazas, lesiones, maltrato, etc.
- **fechaRegistro:** Fecha en la que se registraron los hechos.
- **localidad:** Localidad en la que se produjeron los hechos.
- **observaciones:** Comentarios de la autoridad sobre los hechos.

**Histórico:** Esta tabla contiene información sobre los estados por los que ha pasado el caso. Los campos contenidos son los siguientes:

- **idCaso:** Número identificativo del caso.

**idVictima:** Número identificativo de la víctima.

**idAutor:** Número identificativo del autor.

**fecha:** Fecha en la que se produce el cambio de estado.

**estado:** Cambio que se produce con respecto a la situación anterior. El caso se inactiva o se reactiva.

**motivo:** Motivo por el que se inactiva o reactiva el caso.

**Formulario VPR:** Esta tabla además de recoger las respuestas de los formularios VPR en sí, contiene los siguientes campos:

**idCaso:** Número identificativo del caso.

**idVictima:** Número identificativo de la víctima.

**idAutor:** Número identificativo del autor.

**idVal:** Identificador del formulario VPR.

**tipo:** Tipo de formulario. En nuestra base de datos será siempre VPR 4.0.

**fecha:** Fecha en la que se registra el formulario.

**riskSist:** Riesgo sugerido por el sistema Viogén.

**riskProf:** Riesgo asignado por el agente.

**Formulario VPER:** Esta tabla además de recoger las respuestas de los formularios VPER en sí, contiene los siguientes campos:

**idCaso:** Número identificativo del caso.

**idVictima:** Número identificativo de la víctima.

**idAutor:** Número identificativo del autor.

**idVal:** Identificador del formulario VPER.

**tipo:** Tipo de formulario. En nuestra base de datos puede ser VPER 4.0 con reincidencia o VPER 4.0 sin reincidencia.

**fecha:** Fecha en la que se registra el formulario VPER.

**riskSist:** Riesgo sugerido por el sistema Viogén.

**riskProf:** Riesgo asignado por el agente.

# C

# Detección de reincidencia y delito inicial en formularios

Las respuestas recogidas en los formularios VPER nos indican si ha habido reincidencia y en caso afirmativo, el tipo de reincidencia que se ha producido.

Más concretamente, detectaremos en los formularios VPER que ha habido insultos, si la respuesta a la pregunta F01010 es “Sí” o la respuesta a la pregunta F01011 no es vacía, que ha habido violencia física, si la respuesta a la pregunta F01020 es “Sí” o la respuesta a la pregunta F01021 no es vacía y que ha habido agresión sexual, si la respuesta a la pregunta F01030 es “Sí” o la respuesta a la pregunta F01031 no es vacía. Consideraremos que ha habido violencia, si la respuesta a la pregunta F01000 es “Sí”, la respuesta a la pregunta F01040 es “Sí” o se ha detectado alguno de los subtipos de violencia.

Consideramos que se han utilizado armas contra la víctima, si la respuesta a la pregunta F02000 es “Sí” o la respuesta a la pregunta F02010 no es vacía, que ha habido amenazas, si la respuesta a la pregunta F03000 es “Sí” o alguna de las respuestas a las preguntas F03001 o F03002 está rellena y que ha habido quebrantamientos, si alguna de las respuestas a las preguntas F04000, F04001, F04002 y F04003 es “Sí”.

Las repuestas recogidas en los formularios VPR nos indican el tipo de delito inicial que se ha cometido. Para la detección del tipo de delito inicial a a partir de formularios VPR, se utilizan las respuestas a las mismas preguntas que en el caso de detección de reincidencia en los formularios VPER. La única excepción serán los quebrantamientos, dado que al ser la primera vez que la víctima acude a denunciar al agresor, no es posible que existan y no hay preguntas ligadas a su detección en los formularios VPR.

# D

# Codificación de variables

En este capítulo se muestran las decisiones más importantes que se han tomado a la hora de codificar la base de datos limpia que hemos generado.

## D.1. Codificación de la información contenida en los formularios VPR y VPER

### Codificación de respuestas de formularios

A continuación, se describe la codificación empleada para las respuestas de cada uno de los tipos de pregunta vistos en el apartado 3.2:

**Respuestas a preguntas tipo A:** Se creará una variable, cuyo nombre corresponderá con el identificador de la pregunta en el formulario. Dicha variable tomará el valor 1 si la respuesta es “Sí” y el valor 0 si la respuesta es “No”.

**Respuestas a preguntas tipo B:** Se crearán dos variables binarias cuyo nombre será el resultado de añadir al identificador de la pregunta las terminaciones .S y .N. La variable con terminación .N tomará el valor 1 si la respuesta es “No” y 0 en caso contrario. La variable con terminación .S tomará el valor 1 si la respuesta es “Sí” y 0 en caso contrario. Mediante la concatenación de las dos variables con extensiones .S y .N, las posibles respuestas quedan codificadas como 10 (“Sí”), 01 (“No”), 00(“No sabe”).

**Respuestas a preguntas tipo C:** Se codifican siguiendo el mismo procedimiento que en las respuestas a preguntas de tipo B, con la salvedad de que ahora la codificación 00 corresponderá a la respuesta “No procede” en lugar de a “No sabe”.

**Respuestas a preguntas tipo D:** Se creará una única variable, con nombre el identificador de la pregunta que tomará el valor 0 si no hay respuesta, el valor 1/3 si la respuesta es “Leve”, el valor “2/3” si la respuesta es “Grave” y el valor 1 si la respuesta es “Muy grave”.

**Respuestas a preguntas tipo E:** Por cada una de las posibles respuestas se creará una va-

riable binaria que tomará el valor 1 si la respuesta ha sido seleccionada y 0 en caso contrario. A modo de ejemplo, si una de las posibles respuestas a la pregunta F03002 es MU, se creará la variable binaria *F03002.MU* que tomará el valor 1 si la respuesta ha sido seleccionada y 0 en caso contrario.

**Respuestas a preguntas tipo F:** Se creará una única variable con nombre el identificador de la pregunta, que tomará el valor 0 si la respuesta es “Nulo”, el valor 1/2 si la respuesta es “Bajo” y el valor “1” si la respuesta es “Alto”.

**Respuestas a preguntas tipo G:** : Se crearán dos variables binarias cuyo nombre será el resultado de añadir al identificador de la pregunta las terminaciones .IF y .SB. La variable con terminación .IF tomará el valor 1 si la respuesta es “Infravalora” y 0 en caso contrario. La variable con terminación .SB tomará el valor 1 si la respuesta es “Sobrevalora” y 0 en caso contrario.

### D.1.1. Codificación del delito cometido y del nivel de protección asignado

Para codificar el nivel de protección que asignó el profesional, se creará una variable numérica *riskProfNum*, que tomará el valor 0 si el nivel de protección aplicado es “No apreciado”, 0,25 si es ”Bajo”, 0,5 si es “Medio”, 0,75 si es “Alto” y 1 si es “Extremo”.

Por otro lado, se crearán variables que nos indiquen el tipo de delito que se ha detectado en cada formulario, a partir de las respuestas recogidas en este. En particular, para cada formulario VPR se crearán 7 variable binarias, una por cada tipo de delito inicial presentado en el apartado 3.5. Las variables tomarán el valor 1 si se ha registrado ese tipo de delito a partir de las respuestas del formulario y 0 en caso contrario. Siguiendo el mismo procedimiento, para cada formulario VPER se crearán 9 variables binarias, una por cada uno de los posibles tipos de reincidencia definidos en el apartado 3.5.

## D.2. Codificación de la información general del caso

La información general, que se mantiene invariable a lo largo del caso y va asociada al formulario VPR y a todos los formularios VPER del caso, se codifica de la siguiente manera:

**Edad autor y víctima:** Para codificar la edad del autor se van a dividir las edades en intervalos de longitud 5 y se crearan 15 variables binarias *edadAutor.16-20*, *edadAutor.21-25, edadAutor26-30,..., edadAutor86-90*. La variable *edadAutor.X-Y* toma el valor 1 si la edad del autor está en el intervalo [X,Y] y 0 en caso contrario. La codificación de la edad de la víctima se realiza utilizando el mismo procedimiento.

**Institución:** Para la codificación de la institución se crean cuatro variables binarias: *Institucion.PL, Institucion.PF, Institucion.PN y Institucion.GC*. Cada una de las variables binarias, de la forma *Institucion.X*, toma el valor 1 si el caso ha sido denunciado en la institución X y 0 en caso contrario.

**Localidad:** Las variables *PoblacionLocalidad* y *PropoPoblacionLoc* almacenarán respectivamente, el número de habitantes de la localidad donde se han producido los hechos y su normalización en el intervalo [0,1]. Además se crearán las variables binarias *esPueblo*, *esCiudadPeq*, *esCiudadMed* y *esCiudadGrand*, a partir de las cuales indicaremos la categoría a la que pertenece la localidad. Consideraremos que la localidad es un pueblo si tiene menos de 10.000 habitantes, una ciudad pequeña si tiene entre 10.000 y 59.999 habitantes, una ciudad mediana si tiene entre 60.000 y 125.000, y una ciudad grande si tiene más de 125.000 habitantes.

**Provincia:** con respecto a la provincia se creará la variable binaria *EsFueraPeninsula* que toma el valor 1 si el caso ha sido denunciado fuera de la península y 0 en caso contrario. Además se crearán las variables *PoblacionProvincia* y *PropPoblacionProv* que indican respectivamente, el número de habitantes de la provincia donde se ha denunciado el caso y su equivalente proporcional en el intervalo [0,1]. De manera complementaria, se crearán las variables *esProvPeq*, *esProvMed* y *esProvGrand* que indicarán la categoría de la provincia. Se considera que la provincia es pequeña si tiene menos 350.000 de habitantes, mediana si tiene entre 350.000 y 900.000, y grande si tiene más de 900.000.

## D.3. Codificación del histórico.

### Aumentos y decrementos

Para cada una de las preguntas para las que se recuente si ha habido aumento o decremento en la respuesta con respecto al formulario VPER anterior, se crearán dos variables binarias cuyo nombre será el identificador de la pregunta más las extensiones .A y .D. La variable con extensión .A tomará el valor 1 si se ha producido un aumento y 0 en caso contrario. La variable con extensión .D tomará el valor 1 si se ha producido un decremento y 0 en caso contrario. Se usará la misma técnica para codificar si ha habido aumento o decremento con respecto a cada uno de los tipos posibles de reincidencia.

### Recuento y tendencia de los niveles de protección asignados

Para codificar el número de veces que se ha asignado cada nivel de protección antes del formulario VPER en cuestión, se crean cinco variables: *Acum.NA, Acum.BJ, Acum.MD, Acum.AL* y *Acum.EX*. En cada una de las variables se almacenará el número de veces que se ha asignado con anterioridad el nivel de protección al que hace referencia. Por otro lado, se guardará en las variables numéricas *risk-*

*ProfNum.UltVPER* y *riskProfNum.VPR* el nivel de protección asignado por el profesional en el formulario VPER inmediatamente anterior y en el formulario VPR, respectivamente. Los niveles de protección se guardarán siguiendo la codificación vista en el apartado D.1.1. Para aquellos formularios VPER que sean los primeros del caso, como no tenemos un formulario VPER anterior, se guardará en la variable *riskProfNum.UltVPER*, el nivel de protección asignado por el profesional en el formulario VPR. Para reflejar este último suceso, se creará una variable binaria, *esPrimerVPER*, que toma el valor 1 si es el primer formulario VPER del caso y 0 en caso contrario. La variable *numVPERsPrevios* almacenará el número de formularios VPER registrados anteriormente.

## D.4. Codificación de variables a predecir.

Tal y como se explicó en el apartado 3.7, podemos predecir el NPO para evitar cada uno de los subtipos de reincidencia en el periodo posterior. Por otro lado, podemos predecir simplemente si hay o no reincidencia en el periodo posterior. Crearemos 9 variables, una por cada uno de los tipos de NPO que se puede predecir. Cada una de las variables, tomará el valor 0 si el NPO para evitar el tipo de reincidencia al que esta ligado es no apreciado, 1 si es bajo, 2 si es medio, 3 si es alto y 4 si es extremo. Adicionalmente, crearemos una variable binaria para cada uno de los subtipos de reincidencia a predecir. Cada una de las variables binarias tomará el valor 1 si hay ese tipo de reincidencia en el periodo posterior y 0 en caso contrario.

## D.5. Codificación alternativa

De manera adicional, se propone una versión de codificación alternativa en la que los formularios estén completamente codificados de forma binaria. No obstante, está codificación no ha sido utilizada a la hora de probar los modelos de predicción y se deja como opción para posibles líneas futuras de trabajo.

Para la codificación de preguntas con respuesta tipo D, se crean 4 variables cuyo nombre será el identificador de la pregunta añadiéndole las extensiones “.NL”, “.L” , “.G” y “.MG”. La variable con extensión “.NL” tomará el valor 1 si no hay respuesta y 0 en caso contrario. La variable con extensión “.L” tomará el valor 1 si la respuesta es “Leve” y 0 en caso contrario. La variable con extensión “.G” tomará el valor 1 si la respuesta es “Grave” y 0 en caso contrario. La variable con extensión “.MG” tomará el valor 1 si la respuesta es “Muy grave” y 0 en caso contrario.

Para la codificación de preguntas con respuesta tipo F, se crean 3 variables cuyo nombre será el identificador de la pregunta añadiéndole las extensiones “.NL” , “.BJ” y “.AL” . La variable con extensión “.NL” tomará el valor 1 si la respuesta es “Nulo” y 0 en caso contrario. La variable con extensión “.BJ” tomará el valor 1 si la respuesta es “Bajo” o no hay respuesta y 0 en caso contrario. La variable con

extensión “.AL” tomará el valor 1 si la respuesta es “Alto” y 0 en caso contrario.

Para la codificación del nivel de protección asignado, se crean 5 variables binarias con extensiones “.NA”,”.BJ”,”.MD”,”.AL” y “.EX”, que toman el valor 1 si el nivel de protección asignado es “No apreciado”, “Bajo” ,”Medio”, “Alto” y “Extremo”, respectivamente y 0 en caso contrario.

# E

# Estadísticas sobre los datos.

En este capítulo se muestran algunas de las características extraídas al analizar los datos.

## E.1. Estadísticas sobre la información general

En esta sección queremos ver que relación existe entre las características generales de un caso y la probabilidad de que en ese caso haya reincidencia en algún momento. Por ejemplo, nos gustaría saber como aumenta o disminuye la probabilidad de reincidencia en función de la edad del agresor. Más concretamente, podríamos estudiar como varía esta probabilidad, cuando el agresor tiene entre 26 y 30 años. Para ello, calcularíamos en primer lugar que en un 20,3 % de los casos hay reincidencia. A continuación, seleccionaríamos los casos en los que el agresor tenía entre 26 y 30 años, y calcularíamos en cuantos de ellos ha habido reincidencia. Si este factor no tuviese ninguna dependencia con respecto a la reincidencia, esperaríamos encontrar también dentro de este conjunto que ha habido reincidencia en un 20,3 % de los casos. Si en su defecto encontrásemos que dentro de este conjunto ha habido reincidencia en un 21,3 % de los casos, significaría que la probabilidad se incrementa cuando el caso tiene esa característica.

Vamos a ilustrar como varía la probabilidad de reincidencia en función de cada factor, utilizando varios histogramas. En cada histograma, la altura de la barra indicará el porcentaje de casos en los que se ha dado ese factor. Dentro de la barra, pintaremos en un color diferente el porcentaje de casos en los que hemos obtenido un valor distinto al esperado. Por ejemplo, en la barra que indica en cuantos casos la edad del autor estaba entre 26 y 30 años, si hubiésemos obtenido un 21,3 % de reincidencia en lugar del 20,3 % esperado, pintaríamos en naranja un 1 % de la barra, para indicar el aumento en la probabilidad de reincidencia. Si en lugar de obtener ese resultado, hubiésemos obtenido un 19,3 % de casos con reincidencia, pintaríamos en azul un 1 % de la la barra, para indicar que la probabilidad de reincidencia se reduce.

En la figura E.1 se ilustra la relación entre la edad del autor y la probabilidad de reincidencia. Podemos observar como a medida que aumenta la edad del autor la probabilidad de reincidencia tiende a disminuir.

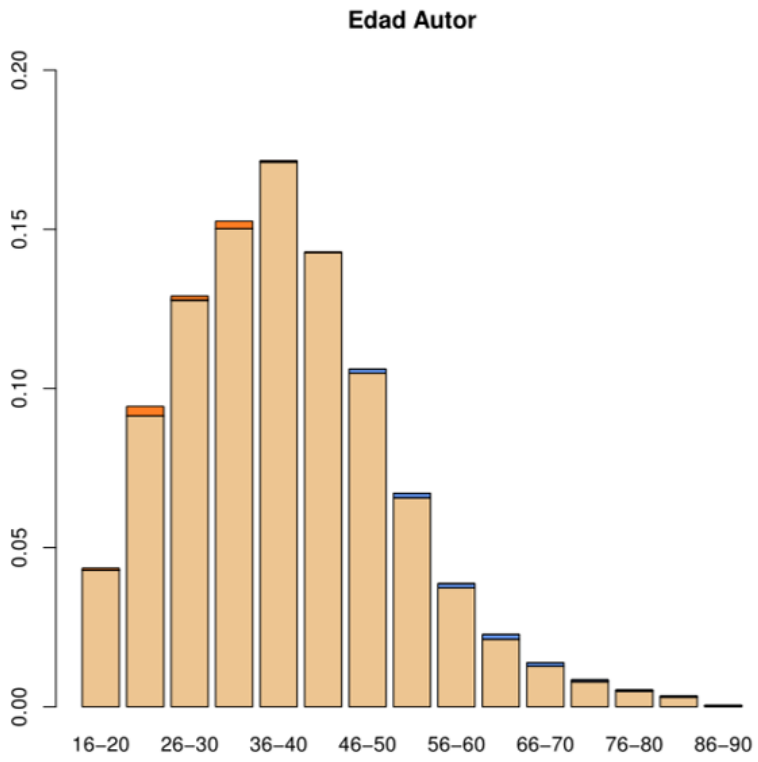


**Figura E.1:** Relación entre la edad del autor y la reincidencia.

De forma análoga, podemos ver a partir de la figura E.2 como a medida que aumenta la edad de la víctima la probabilidad de reincidencia tiende a disminuir. En consecuencia, parece que tanto la edad de la víctima como la edad de la autor, están relacionadas con el riesgo de reincidencia.

Por otro lado, la figura E.3 nos muestra como, a priori, la institución donde se han denunciado los hechos no influye en la probabilidad de reincidencia.

La figuras E.4 y E.5 recogen respectivamente, la relación entre la reincidencia y el tamaño de la localidad y la provincia donde se producen los hechos. Las últimas barras de los histogramas se correspondería con los casos denunciados en Madrid y en la Comunidad de Madrid. Parece a priori que la probabilidad de reincidencia es más baja en esta comunidad. No obstante, hay tantos casos registrados en la Comunidad de Madrid, que se podría hacer otro modelo diferente teniendo en cuenta únicamente estos casos. Para el resto de casos no podemos encontrar una línea de tendencia tan clara como para las edades. Nótese que mediante estos histogramas también se puede estudiar la relevancia de cada variable, una vez la base de datos ha sido codificada.

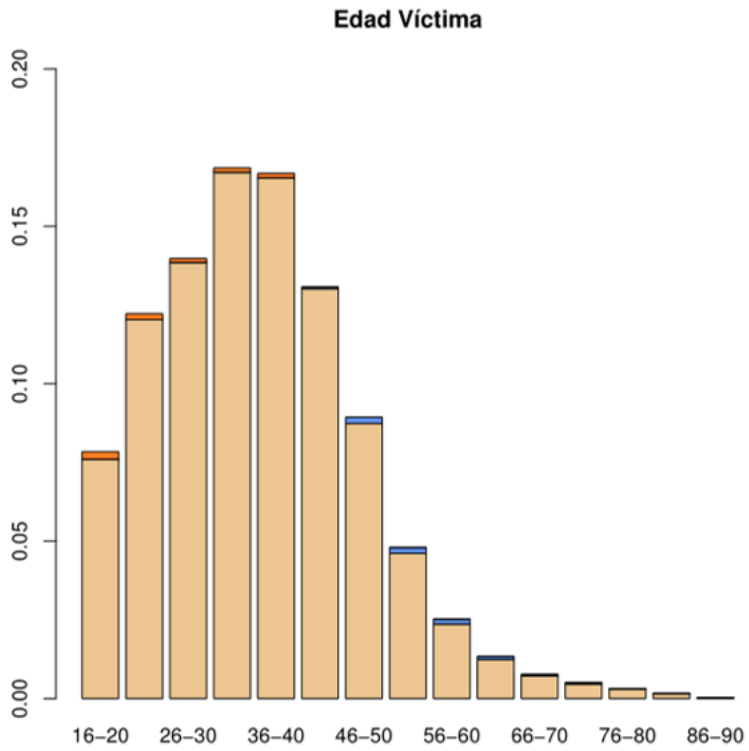


**Figura E.2:** Relación entre la edad de la víctima y la reincidencia.

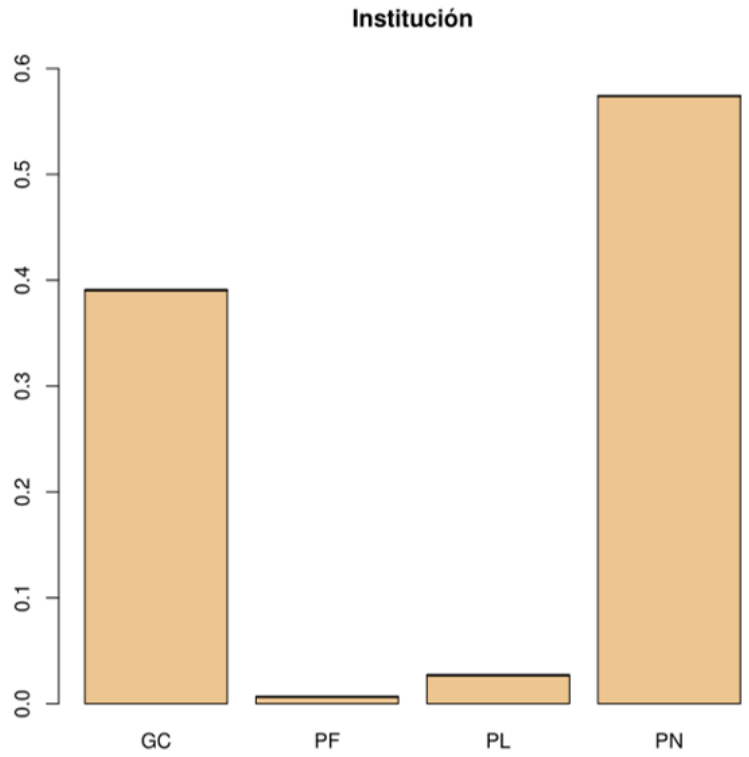


**Figura E.3:** Relación entre la institución y la reincidencia.

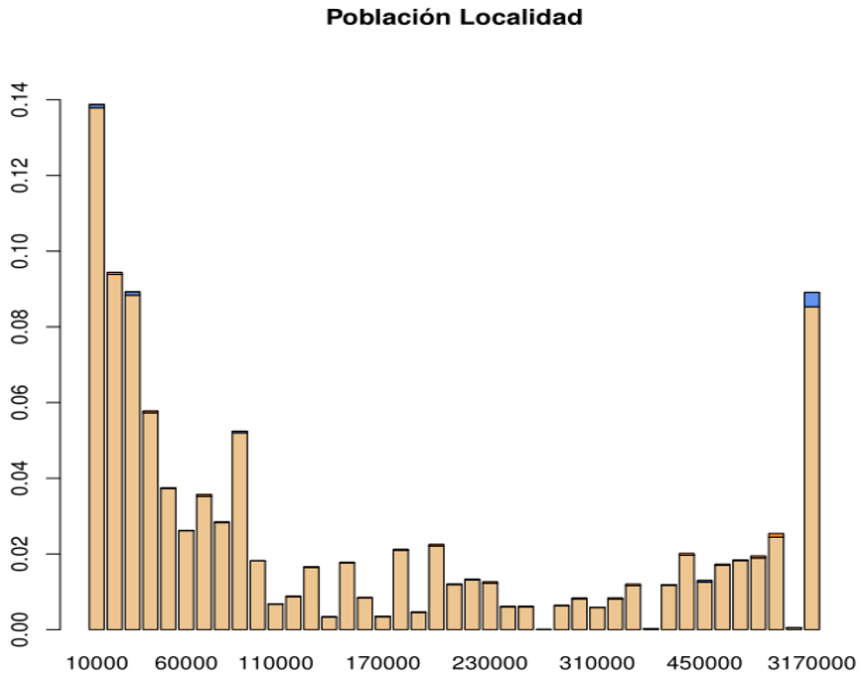


**Figura E.4:** Relación entre el tamaño de la localidad y la reincidencia.

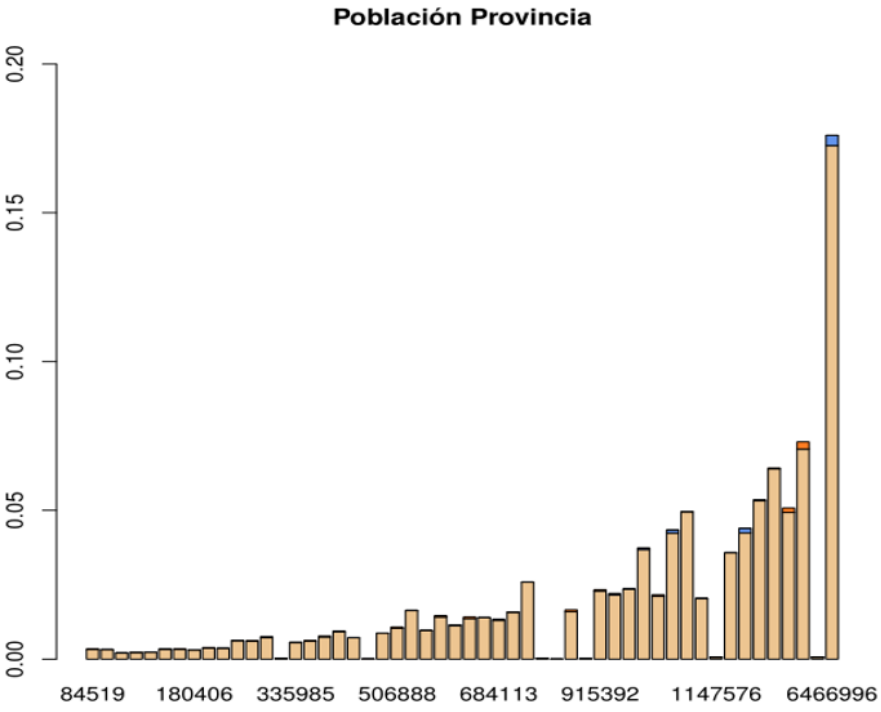


**Figura E.5:** Relación entre el tamaño de la provincia y la reincidencia.

## E.2. Correlaciones de las variables con el nivel de protección óptimo

En la tabla E.1 se muestra la correlación entre el NPO y cada una de las variables del conjunto utilizado para realizar predicciones a partir de formularios VPR. En la tabla E.2 se muestra la correlación entre el NPO y cada una de las variables del conjunto utilizado para realizar predicciones a partir de formularios VPER. El grado de correlación se ha medido a partir del coeficiente de correlación de Pearson en valor absoluto.

| **F03001** | **F12000.N** | **F05000.N** | **F05020.N** | **F05030.N** |
|---|---|---|---|---|
| 0.4742 | 0.4412 | 0.4409 | 0.4345 | 0.4216 |
| **F05010.N** | **F01011** | **F10010.N** | **F06020.N** | **F03000.S** |
| 0.4124 | 0.4116 | 0.3996 | 0.3954 | 0.3946 |
| **F11040.N** | **F09020.N** | **F06040.N** | **F06030.N** | **F11011.N** |
| 0.3924 | 0.3919 | 0.3911 | 0.3891 | 0.3866 |
| **F10020.N** | **F10030.N** | **F11020.N** | **F10070.N** | **F04000.N** |
| 0.3809 | 0.3804 | 0.3772 | 0.3766 | 0.3763 |
| **F09010.N** | **F11010.N** | **F09030.N** | **F10060.N** | **F10090.N** |
| 0.3751 | 0.3695 | 0.3684 | 0.364 | 0.3633 |
| **F06010.N** | **F03000.N** | **F10080.N** | **F07000.N** | **F05020.S** |
| 0.3532 | 0.3525 | 0.3493 | 0.3389 | 0.3255 |
| **F10050.N** | **F09040.N** | **F01010.S** | **F05010.S** | **F05000.S** |
| 0.3244 | 0.3242 | 0.3043 | 0.3033 | 0.3026 |
| **F11030.N** | **F01010.N** | **F03002.MU** | **F10100.N** | **F03002.SU** |
| 0.2951 | 0.2709 | 0.2672 | 0.2666 | 0.2626 |
| **F10040.N** | **F04000.S** | **F05021.FI** | **F10110.N** | **F02000** |
| 0.2547 | 0.2522 | 0.2381 | 0.2289 | 0.218 |
| **F02010.OO** | **F02010.AB** | **F01031** | **F10040.S** | **F01030.S** |
| 0.218 | 0.1999 | 0.188 | 0.1802 | 0.1799 |
| **F05030.S** | **F05021.PS** | **F01021** | **F10110.S** | **F12000.S** |
| 0.1783 | 0.176 | 0.1651 | 0.1589 | 0.1447 |
| **esReinc** | **F01000** | **F06030.S** | **F06040.S** | **F01030.N** |
| 0.1378 | 0.1218 | 0.1048 | 0.1028 | 0.1021 |
| **F02020.N** | **F09020.S** | **F10100.S** | **F01040.S** | **F07000.S** |
| 0.1018 | 0.0992 | 0.0942 | 0.0941 | 0.0824 |
| **F11030.S** | **F06020.S** | **F05021.CB** | **F09030.S** | **F07001.LA** |
| 0.082 | 0.0673 | 0.0632 | 0.0597 | 0.0591 |

| **F08020.N** | **F01020.S** | **F01020.N** | **F08020.S** | **PropoPoblacion-Loc** |
|---|---|---|---|---|
| 0.0579 | 0.0569 | 0.055 | 0.0545 | 0.0539 |
| **F11040.S** | **F01040.N** | **F09040.S** | **F08010.N** | **F09010.S** |
| 0.0517 | 0.0516 | 0.0513 | 0.0504 | 0.0504 |
| **F07001.JU** | **F02010.AF** | **F08010.S** | **PropoPoblacion-Prov** | **F03002.EM** |
| 0.0496 | 0.0471 | 0.0461 | 0.0439 | 0.0434 |
| **F02020.S** | **F08030.N** | **F10050.S** | **F08030.S** | **F02021.DP** |
| 0.0417 | 0.0403 | 0.0395 | 0.0383 | 0.0381 |
| **F06010.S** | **F02021.CZ** | **Institucion.GC** | **Institucion.PN** | **F07001.OS** |
| 0.0368 | 0.0354 | 0.0351 | 0.0335 | 0.0333 |
| **esFueraPeninsula** | **F11010.S** | **F11020.S** | **F05021.ES** | **F10020.S** |
| 0.0318 | 0.0317 | 0.0285 | 0.0272 | 0.0253 |
| **edadVictima.16-20** | **F10060.S** | **F10090.S** | **edadVictima.56-60** | **F10091.PE** |
| 0.0252 | 0.0244 | 0.0237 | 0.023 | 0.0218 |
| **edadAutor.61-65** | **edadAutor.66-70** | **F03002.CH** | **edadVictima.81-85** | **F08000** |
| 0.0201 | 0.0199 | 0.0198 | 0.0195 | 0.0186 |
| **edadAutor.26-30** | **edadAutor.21-25** | **edadVictima.51-55** | **EsPueblo** | **edadAutor.81-85** |
| 0.0181 | 0.0167 | 0.0164 | 0.0161 | 0.0157 |
| **edadAutor.36-40** | **edadAutor.51-55** | **esProvPeq** | **edadVictima.21-25** | **F05021.EC** |
| 0.0145 | 0.0143 | 0.0142 | 0.0141 | 0.0139 |
| **F10080.S** | **edadAutor.76-80** | **edadAutor.41-45** | **edadVictima.71-75** | **edadVictima.26-30** |
| 0.0139 | 0.0139 | 0.0129 | 0.0128 | 0.0118 |
| **F10030.S** | **edadVictima.46-50** | **F03002.RE** | **F10070.S** | **edadAutor.31-35** |
| 0.0112 | 0.0112 | 0.0111 | 0.01 | 0.0099 |
| **edadVictima.61-65** | **edadVictima.31-35** | **edadVictima.76-80** | **edadAutor.56-60** | **edadVictima.66-70** |
| 0.0099 | 0.0097 | 0.0095 | 0.0093 | 0.0088 |

| **edadAutor.86-90** | **edadAutor.46-50** | **F10091.FA** | **edadVictima.41-45** | **esProvMed** |
|---|---|---|---|---|
| 0.0087 | 0.0085 | 0.0084 | 0.007 | 0.0068 |
| **edadVictima.86-90** | **EsCiudadMed** | **edadAutor.71-75** | **edadVictima.36-40** | **F10010.S** |
| 0.0062 | 0.0061 | 0.006 | 0.0058 | 0.0048 |
| **EsCiudadGran** | **F02021.IA** | **F11011.S** | **Institucion.PL** | **EsCiudadPeq** |
| 0.0043 | 0.0041 | 0.0036 | 0.0032 | 0.0027 |
| **edadAutor.16-20** | **esProvGrand** | **Institucion.PF** | | |
| 0.0026 | 0.0017 | 0.0007 | | |

**Tabla E.1:** Correlación con el nivel de protección óptimo de las variables del conjunto asociado a los formularios VPR.

| **riskProfNum.-UltVPER** | **F13000** | **F07020** | **Acum.NA** | **esReinc** |
|---|---|---|---|---|
| 0.62 | 0.418 | 0.4006 | 0.3786 | 0.3541 |
| **F07030** | **F05000.N** | **F07010** | **F05000.S** | **F05030.N** |
| 0.3528 | 0.3501 | 0.3464 | 0.3164 | 0.3153 |
| **Acum.MD** | **F01030.N** | **F09030.N** | **F05020.N** | **F05010.N** |
| 0.3129 | 0.31 | 0.3058 | 0.3034 | 0.2945 |
| **tenMedicina** | **F02000.N** | **F02020.N** | **esReincSin-Queb** | **F08010.N** |
| 0.2939 | 0.2927 | 0.2927 | 0.2887 | 0.2873 |
| **esReinc.A** | **F08010.S** | **F01040.N** | **F08000** | **F04000** |
| 0.2865 | 0.2843 | 0.2837 | 0.2765 | 0.2734 |
| **F09030.S** | **F03001** | **F05020.S** | **F03000.S** | **F01030.D** |
| 0.2677 | 0.264 | 0.2623 | 0.2621 | 0.2583 |
| **F01000.S** | **F05010.S** | **F02000.D** | **F02020.D** | **F09020.N** |
| 0.257 | 0.2515 | 0.2508 | 0.2508 | 0.2463 |
| **F01020.N** | **F01011** | **F01010.S** | **esReinc-SinQueb.A** | **F01040.D** |
| 0.2449 | 0.2445 | 0.2434 | 0.2432 | 0.2415 |
| **F04003** | **F11060.N** | **F08020.N** | **F11070.N** | **F04000.A** |
| 0.2359 | 0.2338 | 0.2337 | 0.2329 | 0.2299 |
| **F07010.D** | **F03001.A** | **F08020.S** | **F09010.N** | **F12060.S** |
| 0.2299 | 0.2278 | 0.2277 | 0.2242 | 0.224 |

| **F01020.D** | **F05000.A** | **riskProfNum.-VPR** | **F10030** | **F10000** |
|---|---|---|---|---|
| 0.224 | 0.2221 | 0.2219 | 0.2214 | 0.2209 |
| **F01010.N** | **F07020.D** | **F04001** | **F07040.S** | **Acum.AL** |
| 0.2151 | 0.2149 | 0.2139 | 0.2126 | 0.2125 |
| **F01021** | **F05030.S** | **F01020.S** | **F01011.A** | **F05030.A** |
| 0.2118 | 0.2108 | 0.2097 | 0.2095 | 0.2081 |
| **F01010.D** | **F07040.N** | **F03002.MU** | **F05020.A** | **F03000.D** |
| 0.2071 | 0.2064 | 0.2063 | 0.2055 | 0.2035 |
| **F01000.A** | **F04003.A** | **F01000.N** | **F05010.A** | **F01000.D** |
| 0.2029 | 0.2022 | 0.2016 | 0.2007 | 0.1993 |
| **F11020.N** | **F03000.N** | **F07060.S** | **F03000.A** | **F12070.S** |
| 0.1989 | 0.1972 | 0.1957 | 0.1926 | 0.1924 |
| **F01010.A** | **F07030.D** | **F11030.N** | **F04001.A** | **F11050.N** |
| 0.1922 | 0.1919 | 0.191 | 0.1874 | 0.187 |
| **F11040.N** | **F13000.A** | **F03002.MU.A** | **F11010.N** | **F01021.A** |
| 0.1828 | 0.1812 | 0.181 | 0.1798 | 0.1798 |
| **F01020.A** | **F05021.FI** | **F12060.N** | **F10020** | **F09020.A** |
| 0.1761 | 0.1756 | 0.175 | 0.1705 | 0.17 |
| **F05021.PS** | **F08030.N** | **F11060.A** | **F09020.S** | **F07040.D** |
| 0.1681 | 0.1667 | 0.162 | 0.1611 | 0.1598 |
| **F11070.A** | **F09010.A** | **F11030.A** | **F11040.A** | **F11010.A** |
| 0.1589 | 0.1582 | 0.156 | 0.1556 | 0.1552 |
| **F09030.A** | **F11020.A** | **F11050.A** | **F08030.S** | **F06000** |
| 0.1543 | 0.154 | 0.153 | 0.1527 | 0.1518 |
| **F01040.A** | **F12070.D** | **F02000.A** | **F02020.A** | **esReinc.D** |
| 0.1493 | 0.1481 | 0.1446 | 0.1446 | 0.1443 |
| **F10010** | **F12080.N** | **F08010.A** | **F01040.S** | **Acum.EX** |
| 0.1441 | 0.1425 | 0.139 | 0.1382 | 0.1382 |
| **F02000.S** | **F02020.S** | **F01030.A** | **F10000.A** | **F11070.S** |
| 0.1377 | 0.1377 | 0.1332 | 0.129 | 0.1286 |
| **F12060.D** | **F09010.S** | **F10030.A** | **numVPER-Previos** | **F12080.S** |
| 0.1273 | 0.1258 | 0.1251 | 0.124 | 0.1237 |
| **F04002** | **F04000.D** | **esReincSin-Queb.D** | **F12070.N** | **F03001.D** |
| 0.1232 | 0.1232 | 0.1221 | 0.1208 | 0.1194 |

| **F02010.OO** | **F08020.A** | **F13001.IF** | **F08000.A** | **F07050.S** |
|---|---|---|---|---|
| 0.118 | 0.1174 | 0.1131 | 0.1131 | 0.1116 |
| **F04003.D** | **F04002.A** | **F05021.FI.A** | **F07060.N** | **F10020.A** |
| 0.1108 | 0.1092 | 0.109 | 0.1075 | 0.1064 |
| **F12080.A** | **F02010.OO.A** | **F01011.D** | **F05021.PS.A** | **F06000.A** |
| 0.1061 | 0.1045 | 0.1038 | 0.1013 | 0.0991 |
| **F10010.A** | **F04001.D** | **F13001.SB** | **F07050.D** | **F03002.MU.D** |
| 0.0983 | 0.0982 | 0.0968 | 0.0965 | 0.0959 |
| **F03002.SU** | **F01021.D** | **F12020.D** | **F08030.A** | **F12050.A** |
| 0.0952 | 0.0947 | 0.0934 | 0.0926 | 0.0899 |
| **F08000.D** | **F05030.D** | **F02010.AB** | **F11060.S** | **F07060.D** |
| 0.0898 | 0.087 | 0.0867 | 0.0862 | 0.0858 |
| **F03002.RE** | **F07010.A** | **F03002.SU.A** | **F05020.D** | **F05021.CB** |
| 0.0856 | 0.0856 | 0.0848 | 0.0844 | 0.0802 |
| **F03002.EM** | **F12030** | **F05010.D** | **F13001.IF.A** | **F02010.AB.A** |
| 0.0794 | 0.0785 | 0.078 | 0.0778 | 0.0769 |
| **F07050.N** | **F05000.D** | **F11050.D** | **F03002.RE.A** | **F11010.D** |
| 0.0766 | 0.0752 | 0.0745 | 0.074 | 0.0738 |
| **F11060.D** | **F01030.S** | **F08030.D** | **F11030.D** | **F11070.D** |
| 0.0729 | 0.072 | 0.0717 | 0.0716 | 0.0707 |
| **F11020.D** | **F11040.D** | **F01031** | **F03002.EM.A** | **F03002.CH** |
| 0.0706 | 0.0696 | 0.0692 | 0.0682 | 0.0669 |
| **F12050.D** | **F09010.D** | **F08020.D** | **F01031.A** | **F05021.CB.A** |
| 0.0666 | 0.0641 | 0.0638 | 0.063 | 0.061 |
| **F12020.A** | **F09020.D** | **F12070.A** | **F11040.S** | **F04002.D** |
| 0.061 | 0.0607 | 0.0607 | 0.0606 | 0.0597 |
| **F07030.A** | **F13001.SB.A** | **F12060.A** | **F02010.OO.D** | **F03002.CH.A** |
| 0.0585 | 0.0585 | 0.0585 | 0.0584 | 0.057 |
| **F08010.D** | **F12040** | **F06000.D** | **Acum.BJ** | **F07020.A** |
| 0.0547 | 0.0536 | 0.0533 | 0.0523 | 0.0503 |
| **F12030.A** | **F07050.A** | **EsPueblo** | **F11050.S** | **F05021.FI.D** |
| 0.0498 | 0.0493 | 0.0491 | 0.049 | 0.0488 |
| **Institucion.GC** | **F12040.A** | **F03002.SU.D** | **Institucion.PN** | **F02010.AB.D** |
| 0.0473 | 0.0468 | 0.0467 | 0.0448 | 0.0434 |
| **F10020.D** | **F10030.D** | **F13001.IF.D** | **F05021.PS.D** | **F10010.D** |
| 0.0432 | 0.0419 | 0.0414 | 0.0412 | 0.0401 |
| **F12010.A** | **F03002.RE.D** | **F10000.D** | **F11020.S** | **F01031.D** |

| | | | | |
|---|---|---|---|---|
| 0.0392 | 0.0386 | 0.0385 | 0.0353 | 0.035 |
| **F03002.EM.D** | **F11030.S** | **F05021.ES** | **F03002.CH.D** | **F11010.S** |
| 0.0336 | 0.0332 | 0.0323 | 0.0318 | 0.0314 |
| **F12080.D** | **edadAutor.71-75** | **edadVictima.56-60** | **F09030.D** | **EsCiudadGran** |
| 0.0301 | 0.0299 | 0.0295 | 0.0284 | 0.0281 |
| **edadVictima.71-75** | **F05021.EC** | **F12050.S** | **F13000.D** | **F05021.CB.D** |
| 0.0277 | 0.0276 | 0.0266 | 0.0266 | 0.026 |
| **edadAutor.76-80** | **F12010.D** | **edadVictima.61-65** | **F07040.A** | **EsCiudadMed** |
| 0.0254 | 0.0254 | 0.0251 | 0.0244 | 0.0228 |
| **F13001.SB.D** | **edadAutor.21-25** | **edadAutor.26-30** | **F05021.ES.A** | **edadAutor.81-85** |
| 0.0221 | 0.0219 | 0.0203 | 0.0194 | 0.0193 |
| **F02010.AF** | **F12010** | **F05021.EC.A** | **edadVictima.26-30** | **esProvPeq** |
| 0.0184 | 0.0184 | 0.0178 | 0.0173 | 0.0172 |
| **F07060.A** | **F02010.AF.A** | **edadVictima.66-70** | **PropoPoblacion-Loc** | **edadVictima.31-35** |
| 0.0169 | 0.0168 | 0.0167 | 0.0166 | 0.0165 |
| **edadAutor.31-35** | **esFuera-Peninsula** | **edadVictima.81-85** | **edadVictima.16-20** | **edadVictima.76-80** |
| 0.0161 | 0.0161 | 0.0158 | 0.0148 | 0.0148 |
| **edadAutor.86-90** | **edadVictima.51-55** | **Institucion.PL** | **edadVictima.46-50** | **Institucion.PF** |
| 0.0143 | 0.0143 | 0.0133 | 0.013 | 0.0123 |
| **F12020** | **edadVictima.86-90** | **esProvGrand** | **F05021.ES.D** | **F02010.AF.D** |
| 0.0104 | 0.0103 | 0.0099 | 0.0095 | 0.0093 |
| **edadVictima.21-25** | **EsCiudadPeq** | **F12030.D** | **F12040.D** | **F05021.EC.D** |
| 0.0091 | 0.009 | 0.0077 | 0.0077 | 0.0074 |
| **edadAutor.16-20** | **edadVictima.36-40** | **PropoPoblacion-Prov** | **edadVictima.41-45** | **esPrimerVPER** |
| 0.005 | 0.0049 | 0.0044 | 0.0025 | 0.002 |
| **F12050.N** | **esProvMed** | | | |

| | | | | |
|---|---|---|---|---|
| 0.001 | 0.0002 | | | |

**Tabla E.2:** Correlación con el nivel de protección óptimo de las variables del conjunto asociado a los formularios VPER .

## E.3. Variables de baja activación

En esta sección se muestran las variables de baja activación. Denominamos variables de baja activación a aquellas que toman el mismo valor en más de un 95 % de los casos. En la tabla E.3 se muestran las variables de baja activación encontradas en el conjunto de variables utilizado para realizar predicciones a partir de formularios VPR. Se incluye el porcentaje de casos en los que cada una de estas variables toma valores distintos al valor más frecuente. De forma análoga, en la tabla E.4 se muestran las variables de baja activación encontradas en el conjunto de variables utilizado para realizar predicciones a partir de formularios VPER.

| | | | | |
|---|---|---|---|---|
| **F06020.S** | **edadVictima.51-55** | **F09040.S** | **F09020.S** | **edadAutor.16-20** |
| 0.0486 | 0.0479 | 0.0458 | 0.0447 | 0.0436 |
| **F06020.S** | **edadVictima.51-55** | **F09040.S** | **F09020.S** | **edadAutor.16-20** |
| 0.0486 | 0.0479 | 0.0458 | 0.0447 | 0.0436 |
| **F02020.S** | **F05021.CB** | **F02010.AB** | **edadAutor.56-60** | **F09010.S** |
| 0.0433 | 0.0417 | 0.0416 | 0.0387 | 0.0287 |
| **Institucion.PL** | **F02021.CZ** | **F10080.S** | **edadVictima.56-60** | **F10060.S** |
| 0.0277 | 0.0276 | 0.0259 | 0.0252 | 0.0233 |
| **edadAutor.61-65** | **F08010.S** | **F10070.S** | **F11011.S** | **F10010.S** |
| 0.0228 | 0.022 | 0.0214 | 0.0196 | 0.0189 |
| **F10020.S** | **F10030.S** | **edadAutor.66-70** | **edadVictima.61-65** | **F07001.JU** |
| 0.0174 | 0.0162 | 0.0138 | 0.0133 | 0.0129 |
| **F02021.IA** | **F10091.FA** | **edadAutor.71-75** | **edadVictima.66-70** | **F05021.EC** |
| 0.0123 | 0.0114 | 0.0085 | 0.0077 | 0.0071 |
| **Institucion.PF** | **edadAutor.76-80** | **edadVictima.71-75** | **F05021.ES** | **F02021.DP** |

| | | | | |
|---|---|---|---|---|
| 0.0069 | 0.0053 | 0.0051 | 0.0048 | 0.0035 |
| **edadAutor.81-85** | **edadVictima.76-80** | **F02010.AF** | **edadVictima.81-85** | **edadAutor.86-90** |
| 0.0034 | 0.0031 | 0.0027 | 0.0017 | 0.0004 |
| **edadVictima.86-90** | | | | |
| 0.0003 | | | | |

**Tabla E.3:** Variables de baja activación en el conjunto de datos asociados a los formularios VPR.

| | | | | |
|---|---|---|---|---|
| **esReinc.D** | **F02000.D** | **F02020.D** | **F01040.D** | **F01030.D** |
| 0.0494 | 0.0489 | 0.0489 | 0.0485 | 0.0473 |
| **Acum.EX** | **edadAutor.16-20** | **edadVictima.51-55** | **F06000** | **esReinc.A** |
| 0.0471 | 0.0466 | 0.0444 | 0.0435 | 0.0424 |
| **F07040.A** | **F10020** | **F07010.A** | **F13001.IF** | **F11070.S** |
| 0.0423 | 0.042 | 0.0381 | 0.036 | 0.0358 |
| **esReincSinQueb** | **F04000** | **F11050.S** | **F10030** | **F07020.A** |
| 0.0356 | 0.035 | 0.0335 | 0.033 | 0.0326 |
| **F05000.D** | **esReincSinQueb-.D** | **F07050.A** | **F04000.D** | **F05030.S** |
| 0.0323 | 0.0302 | 0.0301 | 0.0293 | 0.0288 |
| **F01000.S** | **F07040.D** | **F11010.S** | **F07010.D** | **F04000.A** |
| 0.0283 | 0.0277 | 0.0276 | 0.0271 | 0.0269 |
| **F04003** | **F09030.D** | **F13000.A** | **F07050.D** | **F05021.PS** |
| 0.0267 | 0.0266 | 0.0257 | 0.0253 | 0.0252 |
| **Institucion.PL** | **esReincSinQueb-.A** | **F05020.D** | **F12060.A** | **F05010.D** |
| 0.025 | 0.025 | 0.0247 | 0.0243 | 0.0241 |
| **F12070.A** | **F05030.D** | **F03000.S** | **F03001** | **F12050.A** |
| 0.024 | 0.023 | 0.0229 | 0.0229 | 0.0229 |
| **edadVictima.56-60** | **F01010.S** | **F01011** | **F04003.D** | **F07030.A** |
| 0.0228 | 0.0227 | 0.0227 | 0.0226 | 0.0214 |
| **F04003.A** | **F03001.D** | **F01011.D** | **F07020.D** | **F09020.D** |
| 0.021 | 0.0202 | 0.0199 | 0.0199 | 0.0195 |
| **F11030.S** | **F09030.A** | **F08000.A** | **F05000.A** | **F05021.FI** |
| 0.0193 | 0.0192 | 0.019 | 0.0189 | 0.0187 |

| **F01020.S** | **F01021** | **F07060.A** | **F12080.A** | **F12050.D** |
|---|---|---|---|---|
| 0.0186 | 0.0186 | 0.0186 | 0.0181 | 0.0179 |
| **F12060.D** | **F09010.D** | **F03001.A** | **F07060.D** | **F13001.SB** |
| 0.0176 | 0.0173 | 0.0168 | 0.0168 | 0.0166 |
| **F11040.D** | **F01011.A** | **F01021.D** | **F12080.D** | **F11020.S** |
| 0.0165 | 0.0163 | 0.0159 | 0.0159 | 0.0156 |
| **F08010.A** | **F04001** | **F05020.A** | **F05030.A** | **F11070.D** |
| 0.0156 | 0.0147 | 0.0146 | 0.0146 | 0.0145 |
| **F11060.S** | **F05010.A** | **F11020.D** | **F12070.D** | **F10000.D** |
| 0.0144 | 0.0142 | 0.0133 | 0.0131 | 0.013 |
| **F11060.D** | **F10010** | **F01021.A** | **F04001.D** | **F07030.D** |
| 0.013 | 0.0128 | 0.0128 | 0.0128 | 0.0128 |
| **F08020.A** | **F11050.D** | **F11030.D** | **edadVictima.61-65** | **F03002.MU** |
| 0.0128 | 0.0124 | 0.0118 | 0.0116 | 0.0115 |
| **F04001.A** | **F09020.A** | **F11010.D** | **F12020.A** | **F09010.A** |
| 0.0115 | 0.0115 | 0.0115 | 0.0113 | 0.0109 |
| **F10000.A** | **F03002.MU.D** | **F11040.A** | **F11070.A** | **F10020.D** |
| 0.0104 | 0.0103 | 0.0099 | 0.0099 | 0.0092 |
| **F11050.A** | **F10030.D** | **F03002.MU.A** | **F11020.A** | **F11060.A** |
| 0.0092 | 0.0089 | 0.0086 | 0.0086 | 0.0084 |
| **F08030.A** | **F11010.A** | **F11030.A** | **F05021.PS.D** | **F13001.IF.D** |
| 0.008 | 0.008 | 0.008 | 0.0079 | 0.0079 |
| **F01040.S** | **F10030.A** | **F12020.D** | **F08020.D** | **F08000.D** |
| 0.0078 | 0.0075 | 0.0075 | 0.0075 | 0.0074 |
| **F12030.D** | **edadAutor.71-75** | **F12010.D** | **F05021.CB** | **Institucion.PF** |
| 0.0074 | 0.0073 | 0.0073 | 0.0072 | 0.0072 |
| **edadVictima.66-70** | **F06000.D** | **F10020.A** | **F08030.D** | **F05021.FI.D** |
| 0.0069 | 0.0069 | 0.0068 | 0.0062 | 0.0061 |
| **F13001.IF.A** | **F04002** | **F06000.A** | **F02000.S** | **F02020.S** |
| 0.0059 | 0.0056 | 0.0056 | 0.0055 | 0.0055 |
| **F13001.SB.D** | **F04002.D** | **F04002.A** | **F08010.D** | **edadAutor.76-80** |
| 0.0054 | 0.0049 | 0.0046 | 0.0046 | 0.0045 |

| **edadVictima.71-75** | **F12040.A** | **F03002.RE** | **F12030.A** | **F10010.D** |
|---|---|---|---|---|
| 0.0043 | 0.0042 | 0.0041 | 0.0041 | 0.0041 |
| **F12040.D** | **F05021.PS.A** | **F10010.A** | **F03002.RE.D** | **F05021.FI.A** |
| 0.004 | 0.0039 | 0.0039 | 0.0036 | 0.0035 |
| **F13001.SB.A** | **F12010.A** | **F02010.OO** | **F03002.RE.A** | **F02010.OO.D** |
| 0.0034 | 0.0033 | 0.0032 | 0.0031 | 0.0029 |
| **F03002.EM** | **F05021.CB.D** | **edadVictima.76-80** | **F03002.CH** | **edadAutor.81-85** |
| 0.0028 | 0.0028 | 0.0026 | 0.0026 | 0.0025 |
| **F03002.EM.D** | **F02010.OO.A** | **F03002.CH.D** | **F03002.CH.A** | **F03002.EM.A** |
| 0.0025 | 0.0024 | 0.0022 | 0.002 | 0.002 |
| **F03002.SU** | **F03002.SU.D** | **F05021.CB.A** | **F02010.AB** | **F03002.SU.A** |
| 0.0019 | 0.0017 | 0.0016 | 0.0014 | 0.0014 |
| **F05021.EC** | **F02010.AB.D** | **F01030.S** | **F01031** | **edadVictima.81-85** |
| 0.0013 | 0.0013 | 0.0012 | 0.0012 | 0.0011 |
| **F02010.AB.A** | **F01031.D** | **F05021.ES** | **F01031.A** | **F05021.EC.D** |
| 0.0011 | 0.0011 | 0.0009 | 0.0008 | 0.0004 |
| **edadAutor.86-90** | **F05021.ES.D** | **edadVictima.86-90** | **F05021.EC.A** | **F05021.ES.A** |
| 0.0003 | 0.0003 | 0.0002 | 0.0002 | 0.0002 |
| **F02010.AF** | **F02010.AF.A** | **F02010.AF.D** | | |
| 0.0001 | 0.0001 | 0.0001 | | |

**Tabla E.4:** Variables de baja activación en el conjunto de datos asociado a los formularios VPER.

## E.4. Evolución del nivel de protección

Una de las cuestiones que nos planteamos es como varía el nivel de protección a medida que evoluciona un caso. Para hacernos una idea, calculamos las matrices de transición entre el formulario VPR y el cuarto formulario VPER. Es decir, calculamos cuál es la probabilidad que hay de que la AC asigne cada uno de los niveles de protección, en función del nivel de protección que se asignó en el formulario inmediatamente anterior. A partir de las matrices de transición recogidas en las tablas E.5, E.6, E.7 y E.8, podemos ver como, de manera general, a medida que pasa el tiempo es más probable que se vuelva a asignar el mismo nivel de protección que había vigente.

| Actual\Siguiente | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | **0.6762** | 0.1975 | 0.0973 | 0.0217 | 0.0072 |
| **Bajo** | 0.5737 | **0.3332** | 0.0776 | 0.0120 | 0.0035 |
| **Medio** | 0.3343 | 0.4045 | **0.2408** | 0.0158 | 0.0046 |
| **Alto** | 0.1334 | 0.2785 | 0.4808 | **0.1013** | 0.0059 |
| **Extremo** | 0.0536 | 0.1331 | 0.4351 | 0.3295 | **0.0487** |

**Tabla E.5:** Matriz de transición del nivel de protección entre el formulario VPR y el primer formulario VPER.

| Actual\Siguiente | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | **0.8122** | 0.0918 | 0.0712 | 0.0179 | 0.0068 |
| **Bajo** | 0.3368 | **0.5820** | 0.0646 | 0.0130 | 0.0035 |
| **Medio** | 0.1394 | 0.3672 | **0.4556** | 0.0297 | 0.0082 |
| **Alto** | 0.0466 | 0.1862 | 0.5152 | **0.2429** | 0.0091 |
| **Extremo** | 0.0300 | 0.0950 | 0.3900 | 0.3800 | **0.1050** |

**Tabla E.6:** Matriz de transición del nivel de protección entre el primer formulario VPER y el segundo formulario VPER.

| Actual\Siguiente | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | **0.8879** | 0.0571 | 0.0380 | 0.0130 | 0.0040 |
| **Bajo** | 0.2443 | **0.6857** | 0.0559 | 0.0108 | 0.0033 |
| **Medio** | 0.0975 | 0.3377 | **0.5295** | 0.0268 | 0.0085 |
| **Alto** | 0.0428 | 0.1461 | 0.4924 | **0.3035** | 0.0151 |
| **Extremo** | 0.0110 | 0.1381 | 0.3812 | 0.3812 | **0.0884** |

**Tabla E.7:** Matriz de transición del nivel de protección entre el segundo formulario VPER y el tercer formulario VPER.

| Actual\Siguiente | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | **0.9085** | 0.0508 | 0.0291 | 0.0091 | 0.0026 |
| **Bajo** | 0.1937 | **0.7386** | 0.0522 | 0.0108 | 0.0047 |
| **Medio** | 0.0706 | 0.3145 | **0.5739** | 0.0327 | 0.0083 |
| **Alto** | 0.0232 | 0.1468 | 0.4900 | **0.3200** | 0.0201 |
| **Extremo** | 0.0152 | 0.1439 | 0.3712 | 0.3485 | **0.1212** |

**Tabla E.8:** Matriz de transición del nivel de protección entre el tercer formulario VPER y el cuarto formulario VPER.

## E.5. Variables eliminadas al aplicar Lasso

| | | | | |
|---|---|---|---|---|
| F02010.AF | F02021.DP | F05021.EC | F05021.ES | F10010.S |
| F10060.S | F10091.PE | F10091.FA | edadAutor.66-70 | edadAutor.71-75 |
| edadAutor.76-80 | edadAutor.81-85 | edadAutor.86-90 | Institucion.PF | Institucion.PL |
| edadVictima.61-65 | edadVictima.66-70 | edadVictima.71-75 | edadVictima.76-80 | edadVictima.81-85 |
| edadVictima.86-90 | | | | |

**Tabla E.9:** Variables eliminadas aplicando lasso a M1 para VPR ($\lambda = 0{,}004$)

| | | | | |
|---|---|---|---|---|
| F01011 | F01021 | F01031 | F03001 | F02010.AF |
| F02021.DP | F02021.IA | F03002.RE | F05020.N | F05021.CB |
| F05021.EC | F05021.ES | F05021.FI | F05030.N | F06040.N |
| F07001.JU | F07001.LA | F09040.S | F10010.S | F10010.N |
| F10030.S | F10050.N | F10060.S | F10060.N | F10070.S |
| F10091.FA | F10100.S | F11011.S | esReinc | edadAutor.16-20 |
| edadAutor.56-60 | edadAutor.61-65 | edadAutor.66-70 | edadAutor.71-75 | edadAutor.76-80 |
| edadAutor.81-85 | edadAutor.86-90 | Institucion.PF | Institucion.PL | edadVictima.51-55 |
| edadVictima.56-60 | edadVictima.61-65 | edadVictima.66-70 | edadVictima.71-75 | edadVictima.76-80 |
| edadVictima.81-85 | edadVictima.86-90 | | | |

**Tabla E.10:** Variables eliminadas aplicando Lasso a M2 para VPR ($\lambda = 0{,}005$)

| | | | | |
|---|---|---|---|---|
| F01011 | F01021 | F01031 | F02010.AF | F02021.DP |
| F02021.IA | F05021.EC | F05021.ES | F07001.JU | F10010.S |
| F10030.S | F10070.S | F10090.S | edadAutor.66-70 | edadAutor.71-75 |
| edadAutor.76-80 | edadAutor.81-85 | edadAutor.86-90 | Institucion.PF | edadVictima.61-65 |

| | | | | |
|---|---|---|---|---|
| edadVictima.66-70 | edadVictima.71-75 | edadVictima.76-80 | edadVictima.81-85 | edadVictima.86-90 |

**Tabla E.11:** Variables eliminadas aplicando Lasso a M3 para VPR ($\lambda = 0{,}002$)

| | | | | |
|---|---|---|---|---|
| F05021.EC | F09030.N | F10010.S | F10020.S | F10030.S |
| F10060.S | F10080.S | F11011.S | esReinc | edadAutor.76-80 |
| edadAutor.81-85 | edadAutor.86-90 | Institucion.PF | edadVictima.61-65 | edadVictima.66-70 |
| edadVictima.76-80 | edadVictima.81-85 | edadVictima.86-90 | | |

**Tabla E.12:** Variables eliminadas al aplicar Lasso al submodelo extremo de M4, para VPR ($\lambda = 0{,}008$)

| | | | | |
|---|---|---|---|---|
| F02010.AF | F02021.DP | F05021.EC | F05021.ES | F07001.JU |
| F10010.S | F10020.S | F10091.FA | F11011.S | esReinc |
| edadAutor.66-70 | edadAutor.71-75 | edadAutor.76-80 | edadAutor.81-85 | edadAutor.86-90 |
| Institucion.PF | edadVictima.61-65 | edadVictima.66-70 | edadVictima.71-75 | edadVictima.76-80 |
| edadVictima.81-85 | edadVictima.86-90 | | | |

**Tabla E.13:** Variables eliminadas al aplicar Lasso al submodelo alto de M4, para VPR ($\lambda = 0{,}01$)

| | | | | |
|---|---|---|---|---|
| F01031 | F02010.AF | F02021.DP | F02021.IA | F05021.EC |
| F05021.ES | F07001.JU | F10010.N | F10020.S | F10030.S |
| esReinc | edadAutor.61-65 | edadAutor.66-70 | edadAutor.71-75 | edadAutor.76-80 |
| edadAutor.81-85 | edadAutor.86-90 | Institucion.PF | edadVictima.61-65 | edadVictima.66-70 |
| edadVictima.71-75 | edadVictima.76-80 | edadVictima.76-80 | edadVictima.86-90 | |

**Tabla E.14:** Variables eliminadas al aplicar Lasso al submodelo medio de M4, para VPR ($\lambda = 0{,}005$)

| | | | | |
|---|---|---|---|---|
| F01031 | F02010.AF | F02021.DP | F05021.ES | edadAutor.76-80 |
| edadAutor.81-85 | edadAutor.86-90 | edadVictima.71-75 | edadVictima.76-80 | edadVictima.81-85 |
| edadVictima.86-90 | | | | |

**Tabla E.15:** Variables eliminadas al aplicar Lasso al submodelo bajo de M4, para VPR ($\lambda = 0{,}002$)

| | | | | |
|---|---|---|---|---|
| F01011 | F01021 | F01031 | F03001 | F01030.S |
| F01030.N | F02010.AF | F02010.AB | F02020.S | F02020.N |
| F02021.CZ | F02021.DP | F02021.IA | F03002.SU | F03002.EM |
| F04000.S | F05010.S | F05020.S | F05021.CB | F05021.EC |
| F05021.ES | F05021.FI | F05021.PS | F05030.S | F05030.N |
| F06020.S | F06030.S | F06040.S | F07001.JU | F07001.LA |
| F07001.OS | F08010.S | F08010.N | F09010.S | F09010.N |
| F09020.S | F09020.N | F09030.S | F09040.S | F10010.S |
| F10010.N | F10030.S | F10060.S | F10070.S | F10070.N |
| F10080.S | F10091.FA | F11011.S | F11020.S | F11040.S |
| F12000.S | esReinc | edadAutor.16-20 | edadAutor.56-60 | edadAutor.61-65 |
| edadAutor.66-70 | edadAutor.71-75 | edadAutor.76-80 | edadAutor.81-85 | edadAutor.86-90 |
| Institucion.PF | Institucion.PL | edadVictima.56-60 | edadVictima.61-65 | edadVictima.66-70 |
| edadVictima.71-75 | edadVictima.76-80 | edadVictima.81-85 | edadVictima.86-90 | esProvMed |

**Tabla E.16:** Variables eliminadas al aplicar Lasso al submodelo no apreciado de M4, para VPR ($\lambda = 0{,}001$)

| | | | | |
|---|---|---|---|---|
| edadAutor.71-75 | edadAutor.76-80 | edadAutor.81-85 | edadAutor.86-90 | edadVictima.56-60 |
| edadVictima.61-65 | edadVictima.66-70 | edadVictima.71-75 | edadVictima.76-80 | edadVictima.81-85 |
| edadVictima.86-90 | esProvPeq | F01010.N | F01011 | F01020.N |
| F01021 | F01030.S | F01031 | F01040.S | F02000.S |

| | | | | |
|---|---|---|---|---|
| F02010.AB | F02010.AF | F02010.OO | F02020.S | F03002.CH |
| F03002.EM | F03002.RE | F03002.SU | F04000 | F04002 |
| F04003 | F05021.CB | F05021.EC | F05021.ES | F05021.FI |
| F05021.PS | F05030.S | F11010.S | F11020.S | F11030.S |
| F11040.S | F11050.S | F11060.S | F11070.S | F13001.SB |
| Institucion.PF | Institucion.PN | PropoPoblacion-Loc | esReincSin-Queb.A | F01011.A |
| F01021.A | F01031.A | F02010.AB.A | F02010.AF.A | F02010.OO.A |
| F03001.A | F03002.CH.A | F03002.EM.A | F03002.RE.A | F03002.SU.A |
| F04002.A | F05021.CB.A | F05021.EC.A | F05021.ES.A | F05021.FI.A |
| F05021.PS.A | F06000.A | F07020.A | F07030.A | F10000.A |
| F10010.A | F10020.A | F10030.A | F12010.A | F12020.A |
| F12030.A | F12040.A | F13001.IF.A | F13001.SB.A | esReincSin-Queb.D |
| F01011.D | F01021.D | F01031.D | F02010.AB.D | F02010.AF.D |
| F02010.OO.D | F03001.D | F03002.CH.D | F03002.EM.D | F03002.MU.D |
| F03002.RE.D | F03002.SU.D | F04001.D | F04002.D | F04003.D |
| F05021.CB.D | F05021.EC.D | F05021.ES.D | F05021.FI.D | F05021.PS.D |
| F06000.D | F10000.D | F10010.D | F10020.D | F10030.D |
| F12010.D | F12020.D | F12030.D | F12040.D | F13001.IF.D |
| F13001.SB.D | F05010.D | F05020.A | F05020.D | F05030.D |
| F07050.A | F07060.A | F07060.D | F08010.A | F08010.D |
| F08020.A | F08020.D | F08030.A | F08030.D | F09010.D |
| F09020.D | F09030.D | F11010.D | F11020.A | F11020.D |
| F11030.A | F11030.D | F11040.A | F11040.D | F11050.A |
| F11050.D | F11060.A | F11060.D | F11070.D | F12050.D |
| F12060.A | F12060.D | F12070.A | F12080.D | tenMedicina |

**Tabla E.17:** Variables eliminadas aplicando Lasso a M1, para VPER ($\lambda = 0{,}004$)

| | | | | |
|---|---|---|---|---|
| edadAutor.16-20 | edadAutor.21-25 | edadAutor.71-75 | edadAutor.76-80 | edadAutor.81-85 |
| edadAutor.86-90 | edadVictima.16-20 | edadVictima.31-35 | edadVictima.46-50 | edadVictima.51-55 |
| edadVictima.56-60 | edadVictima.61-65 | edadVictima.66-70 | edadVictima.71-75 | edadVictima.76-80 |
| edadVictima.81-85 | edadVictima.86-90 | esFueraPeninsula | esProvPeq | esReincSinQueb |

| F01000.S | F01010.S | F01011 | F01020.S | F01021 |
|---|---|---|---|---|
| F01030.S | F01031 | F01040.S | F02000.S | F02010.AB |
| F02010.AF | F02010.OO | F02020.S | F03000.S | F03001 |
| F03002.CH | F03002.EM | F03002.MU | F03002.RE | F03002.SU |
| F04000 | F04001 | F04002 | F04003 | F05000.S |
| F05010.N | F05010.S | F05020.N | F05020.S | F05021.CB |
| F05021.EC | F05021.ES | F05021.FI | F05021.PS | F05030.S |
| F06000 | F07030 | F07040.N | F07060.N | F09010.N |
| F09010.S | F09020.N | F09020.S | F09030.S | F10010 |
| F10020 | F10030 | F11010.S | F11020.N | F11020.S |
| F11030.S | F11040.S | F11050.S | F11060.S | F11070.S |
| F12010 | F12030 | F12070.N | F12080.S | F13001.IF |
| F13001.SB | Institucion.PF | Institucion.PN | PropoPoblacion-Loc | esReincSin-Queb.A |
| F01011.A | F01021.A | F01031.A | F02010.AB.A | F02010.AF.A |
| F02010.OO.A | F03001.A | F03002.CH.A | F03002.EM.A | F03002.MU.A |
| F03002.RE.A | F03002.SU.A | F04000.A | F04001.A | F04002.A |
| F04003.A | F05021.CB.A | F05021.EC.A | F05021.ES.A | F05021.FI.A |
| F05021.PS.A | F06000.A | F07020.A | F07030.A | F08000.A |
| F10000.A | F10010.A | F10020.A | F10030.A | F12010.A |
| F12020.A | F12030.A | F12040.A | F13001.IF.A | F13001.SB.A |
| esReincSin-Queb.D | F01011.D | F01021.D | F01031.D | F02010.AB.D |
| F02010.AF.D | F02010.OO.D | F03001.D | F03002.CH.D | F03002.EM.D |
| F03002.MU.D | F03002.RE.D | F03002.SU.D | F04000.D | F04001.D |
| F04002.D | F04003.D | F05021.CB.D | F05021.EC.D | F05021.ES.D |
| F05021.FI.D | F05021.PS.D | F06000.D | F07030.D | F08000.D |
| F10000.D | F10010.D | F10020.D | F10030.D | F12010.D |
| F12020.D | F12030.D | F12040.D | F13000.D | F13001.IF.D |
| F13001.SB.D | F01030.D | F01040.D | F05000.A | F05000.D |
| F05010.A | F05010.D | F05020.A | F05020.D | F05030.D |
| F07040.A | F07050.A | F07050.D | F07060.A | F07060.D |
| F08010.A | F08010.D | F08020.A | F08020.D | F08030.A |
| F08030.D | F09010.A | F09010.D | F09020.A | F09020.D |
| F09030.A | F09030.D | F11010.A | F11010.D | F11020.A |
| F11020.D | F11030.A | F11030.D | F11040.A | F11040.D |
| F11050.A | F11050.D | F11060.A | F11060.D | F11070.A |

| | | | | |
|---|---|---|---|---|
| F11070.D | F12050.A | F12050.D | F12060.A | F12060.D |
| F12070.A | F12070.D | F12080.A | F12080.D | tenMedicina |

**Tabla E.18:** Variables eliminadas aplicando Lasso a M2, para VPER ($\lambda = 0{,}005$)

| | | | | |
|---|---|---|---|---|
| edadAutor.16-20 | edadAutor.21-25 | edadAutor.71-75 | edadAutor.76-80 | edadAutor.81-85 |
| edadAutor.86-90 | edadVictima.16-20 | edadVictima.26-30 | edadVictima.41-45 | edadVictima.46-50 |
| edadVictima.56-60 | edadVictima.61-65 | edadVictima.66-70 | edadVictima.71-75 | edadVictima.76-80 |
| edadVictima.81-85 | edadVictima.86-90 | EsCiudadMed | esProvPeq | EsPueblo |
| esReincSinQueb | F01000.S | F01010.S | F01011 | F01020.S |
| F01021 | F01030.S | F01031 | F01040.S | F02000.S |
| F02010.AB | F02010.AF | F02010.OO | F02020.S | F03000.N |
| F03000.S | F03001 | F03002.CH | F03002.EM | F03002.MU |
| F03002.RE | F03002.SU | F04001 | F04002 | F04003 |
| F05000.S | F05010.N | F05010.S | F05020.S | F05021.CB |
| F05021.EC | F05021.ES | F05021.FI | F05021.PS | F05030.S |
| F06000 | F07030 | F09010.N | F09010.S | F09020.N |
| F09020.S | F09030.N | F10010 | F10020 | F10030 |
| F11010.N | F11010.S | F11020.S | F11030.N | F11030.S |
| F11040.S | F11050.N | F11050.S | F11060.N | F11060.S |
| F11070.S | F12070.N | F12070.S | F13000 | F13001.IF |
| F13001.SB | Institucion.PF | Institucion.PL | Institucion.PN | PropoPoblacion-Loc |
| esReinc.A | esReincSin-Queb.A | F01011.A | F01021.A | F01031.A |
| F02010.AB.A | F02010.AF.A | F02010.OO.A | F03001.A | F03002.CH.A |
| F03002.EM.A | F03002.MU.A | F03002.RE.A | F03002.SU.A | F04000.A |
| F04001.A | F04002.A | F04003.A | F05021.CB.A | F05021.EC.A |
| F05021.ES.A | F05021.FI.A | F05021.PS.A | F06000.A | F07020.A |
| F07030.A | F08000.A | F10000.A | F10010.A | F10020.A |
| F10030.A | F12010.A | F12020.A | F12030.A | F12040.A |
| F13000.A | F13001.IF.A | F13001.SB.A | esReincSin-Queb.D | F01011.D |
| F01021.D | F01031.D | F02010.AB.D | F02010.AF.D | F02010.OO.D |

| | | | | |
|---|---|---|---|---|
| F03001.D | F03002.CH.D | F03002.EM.D | F03002.MU.D | F03002.RE.D |
| F03002.SU.D | F04000.D | F04001.D | F04002.D | F04003.D |
| F05021.CB.D | F05021.EC.D | F05021.ES.D | F05021.FI.D | F05021.PS.D |
| F06000.D | F07010.D | F07020.D | F07030.D | F08000.D |
| F10000.D | F10010.D | F10020.D | F10030.D | F12010.D |
| F12020.D | F12030.D | F12040.D | F13001.IF.D | F13001.SB.D |
| F01030.D | F05000.A | F05000.D | F05010.A | F05010.D |
| F05020.A | F05020.D | F05030.A | F05030.D | F07040.A |
| F07040.D | F07050.D | F07060.A | F07060.D | F08010.A |
| F08010.D | F08020.A | F08020.D | F08030.A | F08030.D |
| F09010.A | F09010.D | F09020.A | F09020.D | F09030.A |
| F09030.D | F11010.A | F11010.D | F11020.A | F11020.D |
| F11030.A | F11030.D | F11040.A | F11040.D | F11050.A |
| F11050.D | F11060.A | F11060.D | F11070.A | F11070.D |
| F12050.A | F12050.D | F12060.A | F12060.D | F12070.A |
| F12070.D | F12080.A | F12080.D | riskProfNum.-UltVPER | tenMedicina |

**Tabla E.19:** Variables eliminadas aplicando Lasso a M3, para VPER ($\lambda = 0{,}003$)

| | | | | |
|---|---|---|---|---|
| edadAutor.71-75 | edadAutor.76-80 | edadAutor.81-85 | edadAutor.86-90 | edadVictima.71-75 |
| edadVictima.81-85 | edadVictima.86-90 | F02010.AF.D | F05021.EC.D | |

**Tabla E.20:** Variables eliminadas al aplicar Lasso al submodelo extremo de M4, para VPER ($\alpha = 0{,}001$)

| | | | | |
|---|---|---|---|---|
| edadAutor.71-75 | edadAutor.76-80 | edadAutor.81-85 | edadAutor.86-90 | edadVictima.56-60 |
| edadVictima.61-65 | edadVictima.66-70 | edadVictima.71-75 | edadVictima.76-80 | edadVictima.81-85 |
| edadVictima.86-90 | F01021 | F01031 | F02010.AB | F02010.AF |
| F05021.EC | F05021.ES | F11050.S | Institucion.PF | F01031.A |
| F02010.AB.A | F02010.AF.A | F02010.OO.A | F03002.CH.A | F03002.EM.A |
| F03002.SU.A | F05021.CB.A | F05021.EC.A | F05021.ES.A | F07010.A |
| F12010.A | F12030.A | F12040.A | F13001.SB.A | F01031.D |

| | | | | |
|---|---|---|---|---|
| F02010.AB.D | F02010.AF.D | F03002.CH.D | F03002.EM.D | F03002.MU.D |
| F03002.RE.D | F03002.SU.D | F05021.CB.D | F05021.EC.D | F05021.ES.D |
| F05021.PS.D | F10010.D | F10020.D | F10030.D | F12010.D |
| F12040.D | F13001.SB.D | F09010.D | F11030.D | F11060.D |

**Tabla E.21:** Variables eliminadas al aplicar Lasso al submodelo alto de M4, para VPER ($\lambda = 0{,}005$)

| | | | | |
|---|---|---|---|---|
| edadAutor.16-20 | edadAutor.21-25 | edadAutor.26-30 | edadAutor.31-35 | edadAutor.71-75 |
| edadAutor.76-80 | edadAutor.81-85 | edadAutor.86-90 | edadVictima.16-20 | edadVictima.21-25 |
| edadVictima.26-30 | edadVictima.31-35 | edadVictima.36-40 | edadVictima.41-45 | edadVictima.46-50 |
| edadVictima.51-55 | edadVictima.56-60 | edadVictima.61-65 | edadVictima.66-70 | edadVictima.71-75 |
| edadVictima.76-80 | edadVictima.81-85 | edadVictima.86-90 | EsCiudadMed | esFueraPeninsula |
| esProvPeq | EsPueblo | esReincSinQueb | F01000.S | F01010.S |
| F01011 | F01020.S | F01021 | F01030.S | F01031 |
| F01040.S | F02000.S | F02010.AB | F02010.AF | F02010.OO |
| F02020.S | F03000.N | F03000.S | F03001 | F03002.CH |
| F03002.EM | F03002.MU | F03002.RE | F03002.SU | F04001 |
| F04002 | F05000.S | F05010.N | F05010.S | F05020.N |
| F05020.S | F05021.CB | F05021.EC | F05021.ES | F05021.FI |
| F05021.PS | F05030.N | F05030.S | F06000 | F07030 |
| F07040.N | F07040.S | F07050.N | F07050.S | F09010.N |
| F09010.S | F09020.N | F09020.S | F10010 | F10020 |
| F10030 | F11010.N | F11010.S | F11020.N | F11020.S |
| F11030.N | F11030.S | F11040.N | F11040.S | F11050.N |
| F11050.S | F11060.N | F11060.S | F11070.N | F11070.S |
| F12010 | F12030 | F12040 | F12070.N | F12080.N |
| F13000 | F13001.IF | F13001.SB | Institucion.PF | Institucion.PL |
| PropoPoblacionLoc | PropoPoblacionProv | esReinc.A | esReincSinQueb.A | F01011.A |
| F01021.A | F01031.A | F02010.AB.A | F02010.AF.A | F02010.OO.A |
| F03001.A | F03002.CH.A | F03002.EM.A | F03002.MU.A | F03002.RE.A |
| F03002.SU.A | F04000.A | F04001.A | F04002.A | F04003.A |
| F05021.CB.A | F05021.EC.A | F05021.ES.A | F05021.FI.A | F05021.PS.A |

| | | | | |
|---|---|---|---|---|
| F06000.A | F07020.A | F07030.A | F08000.A | F10000.A |
| F10010.A | F10020.A | F10030.A | F12010.A | F12020.A |
| F12030.A | F12040.A | F13000.A | F13001.IF.A | F13001.SB.A |
| esReincSin-Queb.D | F01011.D | F01021.D | F01031.D | F02010.AB.D |
| F02010.AF.D | F02010.OO.D | F03001.D | F03002.CH.D | F03002.EM.D |
| F03002.MU.D | F03002.RE.D | F03002.SU.D | F04000.D | F04001.D |
| F04002.D | F04003.D | F05021.CB.D | F05021.EC.D | F05021.ES.D |
| F05021.FI.D | F05021.PS.D | F06000.D | F07010.D | F07020.D |
| F07030.D | F08000.D | F10000.D | F10010.D | F10020.D |
| F10030.D | F12010.D | F12020.D | F12030.D | F12040.D |
| F13000.D | F13001.IF.D | F13001.SB.D | F01030.D | F01040.D |
| F05000.A | F05000.D | F05010.A | F05010.D | F05020.A |
| F05020.D | F05030.A | F05030.D | F07040.A | F07040.D |
| F07050.A | F07050.D | F07060.A | F07060.D | F08010.A |
| F08010.D | F08020.A | F08020.D | F08030.A | F08030.D |
| F09010.A | F09010.D | F09020.A | F09020.D | F09030.A |
| F09030.D | F11010.A | F11010.D | F11020.A | F11020.D |
| F11030.A | F11030.D | F11040.A | F11040.D | F11050.A |
| F11050.D | F11060.A | F11060.D | F11070.A | F11070.D |
| F12050.A | F12050.D | F12060.A | F12060.D | F12070.A |
| F12070.D | F12080.A | F12080.D | riskProfNum.-VPR | esPrimerVPER |
| riskProfNum.-Ult.VPER | | | | |

**Tabla E.22:** Variables eliminadas al aplicar Lasso al submodelo medio de M4, para VPER ($\lambda = 0{,}01$)

| | | | | |
|---|---|---|---|---|
| edadAutor.71-75 | edadAutor.81-85 | edadAutor.86-90 | edadVictima.71-75 | edadVictima.76-80 |
| edadVictima.81-85 | edadVictima.86-90 | F01000.S | F01011 | F01020.S |
| F01021 | F01030.S | F01031 | F01040.S | F02000.S |
| F02010.AB | F02010.AF | F02010.OO | F02020.S | F03001 |
| F03002.CH | F03002.EM | F03002.MU | F03002.RE | F03002.SU |
| F04002 | F05021.EC | F05021.ES | F05021.FI | F01031.A |
| F02010.AB.A | F02010.AF.A | F02010.OO.A | F03002.CH.A | F03002.EM.A |

| | | | | |
|---|---|---|---|---|
| F03002.MU.A | F03002.RE.A | F03002.SU.A | F04000.A | F04001.A |
| F04002.A | F05021.CB.A | F05021.EC.A | F05021.ES.A | F05021.FI.A |
| F05021.PS.A | F06000.A | F10010.A | F10020.A | F10030.A |
| F12010.A | F12030.A | F12040.A | F13001.SB.A | F01031.D |
| F02010.AB.D | F02010.AF.D | F02010.OO.D | F03002.CH.D | F03002.EM.D |
| F03002.SU.D | F05021.CB.D | F05021.EC.D | F05021.ES.D | F06000.D |
| F08000.D | F10010.D | F12020.D | F09010.A | F09020.A |
| F11010.A | F11020.A | F11030.A | F11030.D | F11050.A |
| F11050.D | F11060.A | tenMedicina | | |

**Tabla E.23:** Variables eliminadas al aplicar Lasso al submodelo bajo de M4, para VPER ($\lambda = 0{,}001$)

| | | | | |
|---|---|---|---|---|
| edadAutor.16-20 | edadAutor.21-25 | edadAutor.71-75 | edadAutor.76-80 | edadAutor.81-85 |
| edadAutor.86-90 | edadVictima.16-20 | edadVictima.41-45 | edadVictima.46-50 | edadVictima.56-60 |
| edadVictima.61-65 | edadVictima.66-70 | edadVictima.71-75 | edadVictima.76-80 | edadVictima.81-85 |
| edadVictima.86-90 | EsCiudadMed | esFueraPeninsula | esProvMed | esProvPeq |
| EsPueblo | esReinc | esReincSinQueb | F01000.N | F01000.S |
| F01010.N | F01010.S | F01011 | F01020.N | F01020.S |
| F01021 | F01030.N | F01030.S | F01031 | F01040.N |
| F01040.S | F02000.S | F02010.AB | F02010.AF | F02010.OO |
| F02020.S | F03000.N | F03000.S | F03001 | F03002.CH |
| F03002.EM | F03002.MU | F03002.RE | F03002.SU | F04000 |
| F04001 | F04002 | F04003 | F05000.N | F05000.S |
| F05010.N | F05010.S | F05020.N | F05020.S | F05021.CB |
| F05021.EC | F05021.ES | F05021.FI | F05021.PS | F05030.N |
| F05030.S | F06000 | F07020 | F07030 | F07040.N |
| F07050.N | F07060.N | F09010.N | F09010.S | F09020.N |
| F09020.S | F10000 | F10010 | F10020 | F10030 |
| F11010.N | F11010.S | F11020.N | F11020.S | F11030.N |
| F11030.S | F11040.N | F11040.S | F11050.N | F11050.S |
| F11060.N | F11060.S | F11070.N | F11070.S | F12030 |
| F12040 | F12070.N | F12070.S | F12080.N | F12080.S |
| F13001.IF | F13001.SB | Institucion.PF | Institucion.PL | Institucion.PN |

| PropoPoblacion-Loc | PropoPoblacion-Prov | esReinc.A | esReincSin-Queb.A | F01011.A |
|---|---|---|---|---|
| F01021.A | F01031.A | F02010.AB.A | F02010.AF.A | F02010.OO.A |
| F03001.A | F03002.CH.A | F03002.EM.A | F03002.MU.A | F03002.RE.A |
| F03002.SU.A | F04000.A | F04001.A | F04002.A | F04003.A |
| F05021.CB.A | F05021.EC.A | F05021.ES.A | F05021.FI.A | F05021.PS.A |
| F06000.A | F07010.A | F07020.A | F07030.A | F08000.A |
| F10000.A | F10010.A | F10020.A | F10030.A | F12010.A |
| F12020.A | F12030.A | F12040.A | F13000.A | F13001.IF.A |
| F13001.SB.A | esReincSin-Queb.D | F01011.D | F01021.D | F01031.D |
| F02010.AB.D | F02010.AF.D | F02010.OO.D | F03001.D | F03002.CH.D |
| F03002.EM.D | F03002.MU.D | F03002.RE.D | F03002.SU.D | F04000.D |
| F04001.D | F04002.D | F04003.D | F05021.CB.D | F05021.EC.D |
| F05021.ES.D | F05021.FI.D | F05021.PS.D | F06000.D | F07010.D |
| F07020.D | F07030.D | F08000.D | F10000.D | F10010.D |
| F10020.D | F10030.D | F12010.D | F12020.D | F12030.D |
| F12040.D | F13000.D | F13001.IF.D | F13001.SB.D | F01010.A |
| F01010.D | F01020.A | F01030.D | F01040.D | F02000.D |
| F02020.D | F03000.A | F05000.A | F05000.D | F05010.A |
| F05010.D | F05020.A | F05020.D | F05030.A | F05030.D |
| F07040.A | F07040.D | F07050.A | F07050.D | F07060.A |
| F07060.D | F08010.A | F08010.D | F08020.A | F08020.D |
| F08030.A | F08030.D | F09010.A | F09010.D | F09020.A |
| F09020.D | F09030.A | F09030.D | F11010.A | F11010.D |
| F11020.A | F11020.D | F11030.A | F11030.D | F11040.A |
| F11040.D | F11050.A | F11050.D | F11060.A | F11060.D |
| F11070.A | F11070.D | F12050.A | F12050.D | F12060.A |
| F12060.D | F12070.A | F12070.D | F12080.A | F12080.D |
| esPrimerVPER | Acum.EX | tenMedicina | | |

**Tabla E.24:** Variables eliminadas al aplicar Lasso al submodelo no apreciado de M4, para VPER ($\lambda = 0{,}004$)

# F

# Resultados adicionales sobre los modelos

En este capítulo se muestran los resultados obtenidos al probar los modelos M2, M3 y M4. Para estos tres modelos hemos tenido que eliminar la restricción por la cual se podía infravalorar como mucho un 12,5 % de los casos, puesto que con ningún algoritmo se cumplía.

## F.1. Mejores resultados obtenidos con el modelo M2

Utilizando el modelo M2, el algoritmo que mejores resultados ha generado para la predicción del NPO a partir de formularios VPR, ha sido k-NN con 15 vecinos y métrica *hamming*. La matriz de confusión obtenida se muestra en la tabla F.1. El error obtenido de acuerdo a 4.1 es 0,831.

| NPO real \NPO predicho | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | 8245 | 3107 | 642 | 167 | 6 |
| **Bajo** | 3400 | 6898 | 4828 | 1587 | 95 |
| **Medio** | 395 | 1031 | 4363 | 4240 | 435 |
| **Alto** | 191 | 305 | 689 | 1970 | 679 |
| **Extremo** | 139 | 171 | 239 | 428 | 405 |

**Tabla F.1:** Mejores resultados de predicción del NPO a partir de formularios VPR, empleando el modelo M2

Para la predicción del NPO a partir de formularios VPER, el algoritmo que mejor ha funcionado ha sido Naive Bayes Bernoulli con $\alpha = 0{,}03$ (parámetro de suavizado de Laplace). La matriz de confusión obtenida para este algoritmo se muestra en la tabla F.2. El error obtenido de acuerdo a 4.1 es 0,665.

| NPO real \NPO predicho | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | 83532 | 11481 | 1481 | 639 | 987 |
| **Bajo** | 15987 | 66569 | 8537 | 2530 | 3334 |
| **Medio** | 2192 | 14564 | 13349 | 4369 | 6495 |
| **Alto** | 990 | 2342 | 2956 | 2144 | 3473 |
| **Extremo** | 440 | 733 | 703 | 709 | 2153 |

**Tabla F.2:** Mejores resultados de predicción del NPO a partir de formularios VPER, empleando el modelo M2

## F.2. Mejores resultados obtenidos con el modelo M3

Utilizando el modelo M3, el algoritmo que mejores resultados ha generado para la predicción del NPO a partir de formularios VPR, ha sido Lasso con $\lambda = 0{,}002$. La matriz de confusión obtenida se muestra en la tabla F.3. El error obtenido de acuerdo a 4.1 es 1,358.

| NPO real \NPO predicho | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | 445 | 8068 | 3001 | 583 | 70 |
| **Bajo** | 632 | 10236 | 4734 | 1105 | 101 |
| **Medio** | 252 | 7134 | 2504 | 533 | 41 |
| **Alto** | 47 | 2702 | 876 | 190 | 19 |
| **Extremo** | 21 | 934 | 334 | 86 | 7 |

**Tabla F.3:** Mejores resultados de predicción del NPO a partir de formularios VPR, empleando el modelo M3

Para el caso de predicción del NPO a partir de formularios VPER, el algoritmo que mejor ha funcionado ha sido Naive Bayes Bernoulli con $\alpha = 0{,}03$. La matriz de confusión obtenida al emplear este algoritmo se muestra en la tabla F.4. El error obtenido de acuerdo a 4.1 es 1,240.

| NPO real \NPO predicho | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | 50668 | 23168 | 18542 | 4439 | 1303 |
| **Bajo** | 43301 | 21980 | 20213 | 6880 | 4583 |
| **Medio** | 6800 | 8207 | 11813 | 6351 | 7798 |
| **Alto** | 1236 | 1610 | 2923 | 2088 | 4048 |
| **Extremo** | 436 | 466 | 830 | 725 | 2281 |

**Tabla F.4:** Mejores resultados de predicción del NPO a partir de formularios VPER, empleando el modelo M3

## F.3. Mejores resultados obtenidos con el modelo M4

Utilizando el modelo M4, los algoritmos que han optimizado el valor $F_1$ para cada uno de los submodelos ligado a un nivel de protección y que predice si va a haber reincidencia en la siguiente ventana temporal, son los siguientes:

**Nivel no apreciado:** Naive Bayes Multinomial con $\alpha = 0{,}05$ (parametro de suavización de Laplace).

**Nivel bajo:** Naive Bayes Multinomial con $\alpha = 0{,}1$.

**Nivel medio:** Naive Bayes Multinomial con $\alpha = 0{,}03$.

**Nivel alto:** k-NN con 5 vecinos y métrica *hamming*.

**Nivel extremo:** k-NN con 5 vecinos y métrica *minkowski*.

Los resultados obtenidos para cada uno de los submodelos, cuando se aplican los algoritmos especificados, se muestran en las tablas F.5, F.6, F.7, F.8 y F.9.

| Real/Predicción | No | Sí |
|---|---|---|
| **No** | 4769 | 7398 |
| **Sí** | 792 | 1163 |

**Tabla F.5:** Mejores resultados de predicción de reincidencia en casos en los que se aplica nivel de protección no apreciado, para VPR.

| Real/Predicción | No | Sí |
|---|---|---|
| **No** | 9639 | 5759 |
| **Sí** | 692 | 394 |

**Tabla F.6:** Mejores resultados de predicción de reincidencia en casos en los que se aplica nivel de protección bajo, para VPR.

| Real/Predicción | No | Sí |
|---|---|---|
| **No** | 4132 | 5729 |
| **Sí** | 266 | 356 |

**Tabla F.7:** Mejores resultados de predicción de reincidencia en casos en los que se aplica nivel de protección medio, para VPR.

| Real/Predicción | No | Sí |
|---|---|---|
| **No** | 1278 | 1555 |
| **Sí** | 42 | 74 |

**Tabla F.8:** Mejores resultados de predicción de reincidencia en casos en los que se aplica nivel alto, para VPR.

| Real/Predicción | No | Sí |
|---|---|---|
| **No** | 278 | 300 |
| **Sí** | 16 | 23 |

**Tabla F.9:** Mejores resultados de predicción de reincidencia en casos en los que se aplica nivel de protección extremo, para VPR.

La matriz de confusión obtenida para el modelo M4 cuando se emplean estos algoritmos, se muestra en la tabla F.10. El error obtenido de acuerdo a 4.1 es 2,556.

| NPO real \NPO predicho | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | 331 | 236 | 111 | 1296 | 10193 |
| **Bajo** | 403 | 803 | 373 | 2009 | 13220 |
| **Medio** | 303 | 1136 | 500 | 898 | 7627 |
| **Alto** | 125 | 501 | 228 | 243 | 2737 |
| **Extremo** | 45 | 171 | 63 | 96 | 1007 |

**Tabla F.10:** Mejores resultados de predicción del NPO a partir de formularios VPR, empleando el modelo M4

Para el caso de predicción del NPO a partir de formularios VPER, los algoritmos que mejor resultados han generado para cada uno de los subomodelos han sido los siguientes:

**Nivel no apreciado:** Lasso con $\lambda = 0{,}004$.

**Nivel bajo:** Naive Bayes Multinomial con $\alpha = 0{,}08$.

**Nivel medio:** KNN con 5 vecinos y métrica *hamming*.

**Nivel alto:** Naive Bayes Multinomial con $\alpha = 0{,}01$.

**Nivel extremo:** Lasso con $\lambda = 0{,}001$.

Los resultados obtenidos para cada uno de los submodelos, cuando se aplican los algoritmos especificados, se muestran en las tablas F.11, F.12, F.13, F.14 y F.15.

La matriz de confusión obtenida para el modelo M4, cuando se utiliza estos algoritmos en cada uno de los submodelos, se muestra en la tabla. El error obtenido de acuerdo a 4.1 es 2,629.

| Real/Predicción | No | Sí |
|---|---|---|
| **No** | 39218 | 58902 |
| **Sí** | 4150 | 6257 |

**Tabla F.11:** Mejores resultados de predicción de reincidencia en casos en los que se aplica nivel de protección no apreciado, para VPER.

| Real/Predicción | No | Sí |
|---|---|---|
| **No** | 57473 | 31864 |
| **Sí** | 3059 | 1637 |

**Tabla F.12:** Mejores resultados de predicción de reincidencia en casos en los que se aplica nivel de protección bajo, para VPER.

| Real/Predicción | No | Sí |
|---|---|---|
| **No** | 25146 | 12709 |
| **Sí** | 2234 | 1119 |

**Tabla F.13:** Mejores resultados de predicción de reincidencia en casos en los que se aplica nivel de protección medio, para VPER.

| Real/Predicción | No | Sí |
|---|---|---|
| **No** | 3255 | 3643 |
| **Sí** | 269 | 352 |

**Tabla F.14:** Mejores resultados de predicción de reincidencia en casos en los que se aplica nivel de protección alto, para VPER.

| Real/Predicción | No | Sí |
|---|---|---|
| **No** | 814 | 403 |
| **Sí** | 128 | 57 |

**Tabla F.15:** Mejores resultados de predicción de reincidencia en casos en los que se aplica nivel de protección extremo, para VPER.

| NPO real \NPO predicho | No apreciado | Bajo | Medio | Alto | Extremo |
|---|---|---|---|---|---|
| **No apreciado** | 15680 | 2042 | 1126 | 7864 | 71408 |
| **Bajo** | 14673 | 9560 | 985 | 10092 | 61647 |
| **Medio** | 930 | 8588 | 1039 | 3227 | 27185 |
| **Alto** | 358 | 845 | 307 | 1485 | 8910 |
| **Extremo** | 110 | 219 | 88 | 395 | 3926 |

**Tabla F.16:** Mejores resultados de predicción del NPO a partir de formularios VPER, empleando el modelo M4.